DOCUMENT RESUME

ED 038 051 FL 000 251

AUTHOR Carroll, John B.
TITLE The Prediction of Success in Intensive Foreign Language Training.
PUB DATE 64
NOTE 52p.; Reprint from Training Research and Education, Chapter 4, p 87-136, University of Pittsburgh Press, 1962

EDRS PRICE EDRS Price MF-$0.25 HC-$2.70
DESCRIPTORS *Aptitude Tests, Bibliographies, Diagnostic Tests, Educational Testing, Graphs, *Intensive Language Courses, *Language Ability, Language Instruction, Languages, Learning Difficulties, Predictive Ability (Testing), *Prognostic Tests, Research Reviews (Publications), *Second Language Learning, Statistical Data, Tables (Data), Test Validity

ABSTRACT

After a review of the problem of predicting foreign language success, this booklet describes the development, refinement, and validation of a battery of psychological tests, some involving tape-recorded auditory stimuli, for predicting rate of progress in learning a foreign language. Although the battery was developed for more general application in high schools and colleges, this article focuses on the results obtained in a variety of "intensive" or semi-intensive foreign language courses in a variety of Western and non-Western languages, mainly under governmental auspices, as in the Air Force or at the Foreign Service Institute. Validity coefficients for the tests often reached high levels, with multiple correlations as high as .84. Results suggest there are at least four main components of foreign language aptitude: phonetic coding ability, grammatical sensitivity, rote memory for foreign language materials, and inductive language learning ability. Aptitude is general over different languages, and the tests offer diagnostic possibilities. Results are interpreted in terms of a model that depicts relations among aptitude, ability to understand instruction, motivation, time allowed for learning, and quality of instruction. An appendix of test descriptions and a list of references accompany the text. (Author)

The Prediction of Success In Intensive Foreign Language Training

By John B. Carroll, Harvard University

Reprinted from *Training Research and Education*
University of Pittsburgh Press, 1962

Reprinted December 1964 by Materials Center, Modern Language Association of America, 4 Washington Place, New York, N.Y. 10003, by permission of the author and John Wiley and Sons.

Chapter 4

The Prediction of Success In Intensive Foreign Language Training

John B. Carroll, Harvard University

The advent of World War II brought the military services, and perhaps even the general public, to recognize the desirability of having available, certainly in wartime, considerable numbers of personnel equipped to speak foreign languages of military or political importance. In most cases, this meant that it was necessary to train the requisite personnel almost from scratch, since insufficient numbers had been trained in schools and colleges in critical languages. The "Army" method of intensive language training was developed, largely under the guidance of linguistic scientists, and for the first time in its history, the nation found itself alerted to the possible falsity of the widespread belief that Americans have no aptitude for languages. These "intensive" language learning methods have the drawback of requiring rather large amounts of time. In the program now in effect at the Army Language School (ALS) at the Presidio of Monterey in California, for example, the student devotes almost his entire attention to mastering a foreign language during an eight month, or, in the case of difficult languages, a twelve month period. The length of the training period is similarly long in programs operated by the Air Force, the Navy, the Foreign Service Institute of the Department of State, and other government departments and agencies. No way has been found to reduce the length of the training program beyond a certain point and still produce a satisfactory and useful product. Because of the inevitable expense of these training programs, it is widely accepted that all efforts should be made to minimize the number of training failures both by appropriate screening procedures and by the provision of the best possible instruction. (Training failures are sometimes very numerous. Williams and Leavitt [1947] report an attrition rate of 80% in the Japanese language program in which they did their study.)

Over the past few years the author has been engaged in a program of research on the measurement of aptitude for foreign language learning.[1] Investigations have been made in a variety of settings—in schools, colleges, and universities, in military and governmental training programs, and at the elementary school level as well as at the adolescent and adult level. This chapter is intended to report the major findings that are applicable to the screening of personnel for military and governmental programs for intensive or semi-intensive foreign language training.

A word should be said, however, about research in foreign language training methods. The word can be brief, however, because the amount of psychological research on language training methods in military settings has been pitifully small. One brief attempt was made to carry out a program of such research, but this had to do with the teaching of English to foreigners. A series of instructional materials were developed, but the program was terminated before these could be properly evaluated. One or two research papers on fundamental questions in language training were published (e.g., Kopstein & Roshal, 1954), but they could not begin to answer all the interesting and pertinent questions that could be raised about the psychology of language training. Elsewhere, the psychological research literature in this area has been reviewed (Carroll, 1962), but very little of this research has been done in the setting of a modern, intensive language training program emphasizing the attainment of fluent speaking and reading comprehension of the foreign language. Much research remains to be done; some of the funds appropriated under Title VI of the National Defense Education Act of 1958 have become available for psychological research on methods of language teaching. Up to now, the major need has been for methods that are better organized from a linguistic point of view—texts, tapes, and films that present language sounds, patterns and structure in a clear and well-ordered way, based on adequate linguistic re-

[1]This research was supported chiefly by grants to Harvard University from the Carnegie Corporation of New York. That Corporation, however, is not to be held responsible for opinions expressed here, which are solely the responsibility of the author. Thanks are due to all who were associated with this project, particularly to Dr. Stanley M. Sapon of Ohio State University who spent two years (1953-1955) with the author in developing and studying foreign language aptitude measures. The author is indebted to Robert Gardner for helpful comments regarding a preliminary version of this chapter.

search. Even with the best linguistic research, however, there still remain questions which one would hope psychological research could answer, such as: How much will use of the native language help or hinder learning? How long or how often should a given item be practiced? How can the teaching of grammar be organized to give the student maximum flexibility in varying language patterns?

The theme here, however, is selection, and at the outset two propositions should be mentioned, the truth of which this chapter will attempt to demonstrate, and which if accepted will serve to accent the needs which prompted the program of research reported. These propositions are (*a*) that facility in learning to speak and understand a foreign language is a fairly specialized talent (or group of talents),[2] relatively independent of those traits ordinarily included under "intelligence" and (*b*) that a relatively small fraction of the general population seems to have enough of this talent to be worth subjecting to the rigorous, intensive, expensive training programs in foreign languages operated by military and governmental organizations, or by such private organizations as missionary societies, businesses, and industries engaged in overseas operations. This latter proposition is not meant to imply that the American population is abnormally low in foreign language aptitude, for, in any case, there are no comparative data from non-American populations. Further, it should not be taken to suggest that only a relatively small number of American school children or college students ought to study foreign languages. The question of whether a student of lower than average aptitude should study foreign languages for purposes of general and liberal education depends upon a number of important considerations which do not bear upon the selection of students for intensive foreign language courses of the type described here. The proposition applies only to situations where the organization sponsoring a language training course has a legitimate reason for exercising selectivity.

---

[2]In speaking of "talent" and "abilities," the conveniences of ordinary language are being indulged in; from a strictly operational behavioristic point of view the references would have to be to correspondences observed between behavior on tests and behavior in training programs, with possibly some inferences about "prior learning," "learning sets," etc. It still remains to be investigated to what extent the behavior measured on the aptitude tests which were developed can be modified by training and whether such learning would transfer to real language learning situations.

The proposition that foreign language aptitude is relatively specialized can be introduced by pointing out the well-known fact that intelligence tests have been largely unsuccessful in screening individuals for language training. To be sure, certain cutting points may be introduced to eliminate those of limited intellectual ability, but there are apparently wide variations in the language learning ability of those who are above the cutting point. The screening procedure used by the U. S. Air Force before the advent of valid aptitude tests may be cited. As described by Frith (1953), this involved the use of a so-called "trial course" in a foreign language, either in Russian or in Chinese; success in the trial course, of one to three weeks' duration, was required before the individual was selected for longer, intensive language training. Requirements for entrance into the trial course were: (*a*) an Armed Force Qualification Test (AFQT) score of 70 or better; (*b*) a Technician Specialty Aptitude Index of eight or better;[3] (*c*) a high school diploma or the equivalent; and (*d*) a desire to study the language. Despite the imposition of requirements with an intellectual component, success in the trial course varied widely. In fact—and this supports the second proposition—the proportion of those selected for further, intensive training was relatively small, sometimes as low as 25%. In two trial courses in Chinese in which the language aptitude tests which were developed were tried out, the figures were 41% and 44%, based only on those who did not *voluntarily* withdraw from the course before its completion. The figures were 30% and 38% when based on the total input. These low percentages were not the result of small quotas and undue selectivity; they were the result of experience which showed that it was necessary to enter large numbers of personnel into the trial course in order to obtain candidates with what appeared to be the requisite aptitude. In fact, in the tryouts of aptitude tests, the test results agreed well with the appraisals of the instructors who ran the trial course; in two such trial courses the multiple correlations were .76 and .84 respectively. Furthermore, the evidence avail-

[3]The following explanation of these scores was furnished by Dr. Francis D. Harding of the Personnel Laboratory, Wright Air Development Center (Air Research and Development Command), Lackland Air Force Base, Texas: "The scores on the AFQT represent percentile scores in a sample equivalent to the World War II mobilization population. The Technician Specialty Index of eight converts to about a 75 percentile score. The Technician Specialty Index, now known as the General Aptitude Index, is really a measure of general ability."

able suggests that both the aptitude tests and the instructors' appraisals in the trial course were valid for predicting success in further training.

Later in the chapter the details of the experiments alluded to above will be reported but it seems now sufficiently well established that there was need for better selective devices than intelligence tests, trial courses, and the like.

Previous Research

If intelligence tests were not adequate for screening candidates for language training, why was there no resort to previously published language aptitude tests? The reasons for this are complex, and must be presented in the light of a brief history of language aptitude testing.

Apparently the first efforts to develop aptitude tests for foreign language study were made in the second and third decades of this century. This work was summarized in a publication edited by Henmon (1929). The tests were generally of two sorts: (*a*) tests of ability and achievement in the English language—vocabulary, grammar, spelling, etc., and (*b*) work-sample testing involving short "lessons" or problems in the language to be studied or in an artificial language. In every case, however, the tests were of the paper-and-pencil variety and emphasized ability to deal with the intellectual, cognitive aspects of language study, that is, in the main, with the learning of a written language. In the 1920's and 1930's, for various reasons which will not be developed here, the main objective of foreign language study in the schools was to teach the student to read, or perhaps it would be stated better to say—to translate a foreign language. Some of the tests developed during this period, e.g., the Iowa Foreign Language Aptitude Examination (Stoddard & VanderBeke, 1925), the Symonds Foreign Language Prognosis Test (Symonds, 1930), the Luria-Orleans Modern Language Prognosis Test (Luria & Orleans, 1928), and the George Washington University Language Aptitude Test (Hunt, Wallace, Doran, Buynitzky, & Schwarz, 1929) were reasonably effective in predicting success in language learning under these conditions. However, they tended to have high correlations with intelligence tests; indeed, one part of the American Council on Education Psychological Examination is an artificial language test presenting vocabulary and grammar rules in an artificial language which the student must apply in completing the test. They presented an essentially intellectual task which the student could solve by an analytical procedure quite similar to that which

was demanded by the kind of foreign language teaching which tended to be prevalent at the time. But since success in these courses could generally be predicted about as well by intelligence tests as by the special prognostic tests, the latter were not widely used. Another characteristic of these tests which undoubtedly affected their validity (either increasing or decreasing it, depending on the nature of the criterion) was that they assumed or tested certain specific prior learnings, such as the knowledge of grammatical terminology, and recognition of morphological processes like prefixing and suffixing.

During World War II, selection of men for training in intensive language courses was based mainly on amount of education. The Army developed a work-sample language aptitude test, but did so too late for wide use during the war. It was pressed into service, however, as one part of a battery of tests for aiding in the selection of candidates for the U. S. Military Academy at West Point (for this reason the test appears sometimes in the literature with the designation WPAT or WPQ). This test is still in use for that purpose; data on its validity in one West Point class that was studied will be given below. Even so, the test was not designed with the requirements of modern intensive language training in mind and continued the tradition of posing linguistic puzzles in an artificial language that could be solved analytically.

There were sporadic attempts to investigate the prediction of language learning success in intensive language learning contexts. Bottke and Milligan (1945) suggested several types of test items which might bear on aural and oral abilities, but they did not publish results of any kind. Williams and Leavitt (1947) investigated the usefulness of a series of tests in predicting success or failure in an intensive course in spoken and written Japanese, conducted by the U. S. Marine Corps during World War II. The tests included the U. S. Navy Officer Qualification Test (with three parts: Verbal Opposites, Mechanical Comprehension, and Arithmetical Reasoning); the U. S. Army Language Aptitude Test WPQ-1; the American Council on Education Psychological Examination for college freshmen (with separate scores on the Language and Quantitative subtests); Anderson's Adult Placement Test (parts 1, 2, 3, only); the Shipley-Hartford Retreat Scale; the National Defense Research Council (NDRC) Personal Inventory Form; a "specially devised Symbol Digit Test in which the symbols were Japanese-like nonsense characters"; a "specially devised Figure Recognition Test of visual memory of geometric forms"; the AGCT (Army General Classification Test); and the Army Mechanical Aptitude Test.

Critical ratios greater than 3.0 between means of 27 passers and 48 failers were obtained only for the ACE Language Score, the Army Language Aptitude Test, the Verbal Opposites test of the Navy Officer Qualification Test, and the Vocabulary section of the Shipley-Hartford Retreat Scale—all verbal tests. Computations from data presented by these authors yield biserial *r*'s of .71 for the ACE language score and .70 for the Army Language Aptitude Test. Williams and Leavitt were evidently working in a situation where verbal intelligence was a good predictor of success in intensive foreign language training. It would be possible to organize and teach a course in such a way that verbal intelligence would be at a premium, and it is conceivable that this was true for the Marine Corps course. The studies to be reported here, however, demonstrate that verbal intelligence will not always be a good predictor. This will be seen in the fact that the Cooperative Vocabulary Test is only a moderately good predictor in Tables 4.1 and 4.10.

The first large-scale study of foreign language aptitude after World War II was conducted by Dorcus, Mount and Jones (1952) under a contract with the Department of the Army. Working at the Army Language School in California, they investigated two sets of variables as possible predictors of language success. The first set of variables comprised data already available for 279 cases from files of the school, including scores on the AGCT, three subtests of the Seashore Musical Aptitude Test (Tonal Memory, Timbre, and Pitch), and the Army Language Aptitude Test WPQ-1. Of these variables, only the last yielded any significant correlation with language grades, and then chiefly with grades in the first two weeks of the course. The second set of variables, applied to 152 new cases, included a series of ten novel, specially-constructed tests of abilities in the verbal, auditory, perceptual, and memory domains; six of these were presented by means of magnetic tape. The new variables failed to produce any impressive gain in predictive power over the Army Language Aptitude Test WPQ-1; in fact, in most cases they themselves had non-significant validity coefficients.

In appraising this study, one can hardly criticize the criterion variables, on which much labor was expended. Measures of both spoken and written proficiency for six languages taught at the ALS were constructed, and grades were obtained for both initial and final phases of the course. The authors argue, in fact, that the high correlations of early and final course examinations "indicate that important factors do exist for the prediction of language proficiency . . . on the assumption that

insufficient learning has occurred in the first few weeks of training to account for the relationship to the final examination score" (Dorcus, et al., 1952, p. 3).

It may be suggested that the Dorcus, Mount and Jones study failed to achieve satisfactorily high predictive validities for the new tests because these tests just missed measuring certain abilities that were crucial in language learning. The tasks sampled were not sufficiently close, in behavioral structure, to the tasks actually involved in learning a new language. For example, language learning does not involve memory span for *digits* (as was measured in one of the tests), though it might well involve memory span for *speech sounds*; likewise, there is nothing in language learning which resembles the "Difficult Reading" task (e.g., stating how many English words are found in the following sequence: wo Uld HaRd;ly bef At Alev entho Ugh hew As by hiMse;lf and).

The wide variations in the success achieved by various projects seeking to predict progress in learning foreign languages further indicated that there was an acute need not only for tests which would reliably predict success in different kinds of foreign language courses, but also for better knowledge of the factors making for success. Tests would be useful not only in selection, but also in guidance, placement, and research. Knowledge about factors making for success might eventually make it possible to improve teaching procedures so as to overcome some of the commoner student difficulties. Idealistically, one would like to see instruction improved to the point where the need for aptitude tests will be eliminated, but that day does not seem to be very near at hand. In any case, valid language aptitude tests would be highly useful for providing controls in experimentation on language teaching methods. It appeared, then, that an urgent practical problem was to be faced, as well as an interesting scientific puzzle.

Test Development

In embarking on the research it seemed that there was a place for broad-scale empiricism, guided when possible by theory, and where that was impossible, by hunches. Random experimentation in prediction studies is to be deplored, to be sure (Travers, 1954; Travers, 1956) and it was seldom the case, in these studies, that a predictor variable was investigated simply because it was available. The primary consideration in selecting and devising the tests of the initial trial batteries was to include a variety of tests each of which promised to measure

some aspect of the complex of traits deemed requisite for success in the criterion performance.

The first battery which was tried out contained 20 separate tests and included additional predictor variables that were obtained in several of the samples used for the tryouts. In assembling the tests, one of the guiding principles was to include tests of several of the established "factors" or dimensions of the domain of verbal abilities. It seemed reasonable (Carroll, 1953) to suppose that some of the dimensions of individual differences observable through tests of ability with the English language might also manifest themselves i learning a second language. The factors of verbal ability thus included were: (*a*) the verbal knowledge factor V; (*b*) the word-fluency factor *W* (which can be regarded as mainly involving knowledge of orthographic habits); (*c*) the fluency-of-expression factor *FE*; (*d*) the associative memory factor *M*; and (*e*) the naming factor *Na*. (The designations given by French [1951] are also used here.) In the case of each of these factors, it was hypothesized that the behaviors measured by the corresponding tests had certain elements in common with behaviors involved in foreign language learning. These hypotheses are, of course, incapable of direct confirmation, although they would tend to gain support if one were to obtain consistently significant positive validity coefficients for the corresponding tests.

Other tests were developed and included in the initial tryout battery because they were believed to measure certain specific abilities required in second language learning. One was essentially a "grammatical analogies" test in which the terms of the analogy were relations between a given linguistic form, a word or phrase, and the sentence in which it is placed; thus, in the following sample item

The man went into the HOUSE.

The *church* next to the *bowling alley* will be built in a new *location* next *year*.
A B C D E

the examinee has to find which lettered element in the second sentence has the same relation to its sentence as "house" has to its sentence. The test therefore appears to measure "grammatical sensitivity," that is, the ability to recognize the grammatical function of words in sentences ("Words in Sentences" is the name for this test), and it does this without at the same time requiring the examinee to know the meaning of such grammatical terms as noun, adjective, predicate, preposition and the like. It was thought very likely that some kind of

sentence-analysis would be involved in foreign language learning, regardless of whether the emphasis was on speech or writing.

Still another test included in the battery to fit a particular hypothesis was a test of Phonetic Discrimination which was developed by Dr. Stanley M. Sapon. One of the first tasks which the learner of a foreign language has to master is to recognize the differences among the sounds of the foreign language; often these sounds appear to be so similar to the native speaker of English as to be almost indistinguishable, that is, while they are phonemically distinct in the foreign language, they are not so in English. It was therefore believed that a test of the ability to perceive phonetic distinctions would be a useful item in an aptitude battery. In the initial form of the test, a series of triads of sounds in various languages were presented by tape recording; in each triad, two of the sounds were alike and the third was different, the subject being instructed to identify the odd member. This test, incidentally, tended to be too easy and to have low reliability; a later form used the more conventional multiple-choice type of item, but this was still unreliable. What is more, the validity coefficients were consistently low in comparison to those of other tests, and the conclusion was reached that phonetic discrimination ability is not crucial in foreign language learning. Most normal people have enough discrimination ability to serve them in learning a foreign language, and in any case, it is more a matter of *learning* the discrimination over a period of time than any fundamental lack of auditory discrimination which can readily be tested in an aptitude battery.

One other consideration in the assembly of the initial tryout battery was the desirability of including some "work-sample" tests. The use of work-sample tests in aptitude batteries has sometimes been criticized, either on account of the low level of theoretical sophistication that they imply, or on account of the high specificity which they seem to involve. It is claimed that the too frequent use of work-sample tests may imply that one needs a different work-sample test for every conceivable kind of task for which one might want to predict an individual's success. These arguments do not apply in the present case, and they also mistake the potential value of the work-sample test. To be sure, it is perhaps regrettable that a special language prognosis test is needed to supplement other kinds of tests in the psychometrician's kit, but as long as it is—presumably because of the specific nature of language aptitude—there is no reason to exclude a work-sample test if such a test is found to have adequate validity. (On work-sample

tests, see Chapter 12 by Wilson.) Furthermore, a work-sample test is seldom as specific as has been claimed. As a sample of the kinds of tasks to be learned, it may very well require the same abilities that are required in a broad class of criterion tasks. The abilities may appear to be specific to this class of criterion tasks only because nobody has invented a way of tapping them in other connections.

Several kinds of work-sample tests were devised in the course of this project. One of the simplest was adapted from a test developed some years before for research on verbal abilities (Carroll, 1941). This was an artificial language test in which the examinee has to learn the names for a simple foreign language number system, after which he is required to write down three-digit numbers from dictation. The examinee must not only learn the number system; he also has to be able to attend to and respond to a complex auditory signal, i.e., the artificial language numbers as read from a tape recording. The behavior can be regarded as highly similar to that of understanding a foreign language when spoken rapidly.

Another work-sample test, tried out in the initial battery, was one in which an attempt was made to simulate modern oral language learning as closely as possible. This test, described by Sapon (1955), was perhaps one of the first fully "automated" language teaching sequences; it presented the lesson material in easy steps by a tape recording synchronized with a film strip. There was no use of English beyond initial directions, and the lesson material was arranged so as to make possible inductive learning of the vocabulary and grammar of "Tem-Tem," the artificial language constructed for this test.

Still another work-sample test was devised to simulate *traditional* language instruction of the grammar-translation type designed with solely the reading objective in mind. The test was designed to outdo the Army's West Point Aptitude Test by being based more squarely on linguistic principles. It, too, was administered by tape, but traditional grammar lessons were read and explained to the examinee, with proper attention to the pronunciation of words in the artificial language, "Perdašeb," contrived for the purpose of the test.

Work samples such as these have the drawbacks of complexity and of excessive time requirements. Both the Tem-Tem and Perdašeb tests required nearly a half-hour of testing time apiece, and the former required the use of a colored strip-film in a suitable projector. They could not be used in the final test battery which resulted from the

project, but they were valuable in the research program for helping to delineate the nature of language aptitude.

The project was conceived of, in short, as an opportunity to try out different kinds of tests and modify hypotheses until more was known about factors involved in language aptitude. At a rather early stage in the project, two correlation matrices were subjected to factor-analysis study (Carroll, 1958). It is probably more appropriate to defer discussion of these results and their implications for the nature of foreign language aptitude until after presenting results of the several validation studies that were conducted. The following sections provide an overview of these results. In general, only results pertinent to intensive language training will be included. Validation studies performed in connection with regular academic courses will be presented elsewhere. Results for two service academies, the U. S. Military Academy at West Point and the U. S. Air Force Academy in Colorado, will be included here because their courses may be described as semi-intensive.

Test Validity in Intensive Language Courses

The paramount concern in any research on aptitude is the validity of the test. This and several following sections will provide an overview of those phases of the research program which were designed to identify the most promising tests of language aptitude and to investigate the conditions under which they were valid.

At the outset, it was desirable to try out a large number of tests in a situation where criterion measurements would not be long in becoming available. The tests were tried out first in connection with one of the "trial courses" conducted by the U. S. Air Force to screen personnel for further language study. In February 1954, a battery of 20 tests, totalling about four hours of testing time, was administered to 111 men who had been previously screened by the requirements listed above (Frith, 1953) and who had volunteered to try for the chance of being selected for an eight-months' course in Mandarin Chinese at the Institute of Far Eastern Languages (IFEL), Yale University. The list of tests is given in Table 4.1. (All of the tests mentioned here are briefly described in the Appendix at the end of the chapter.) Immediately after the completion of testing on Monday, the men started intensive study of spoken Mandarin Chinese under experienced instructors, and lessons continued through Friday. Of the 111 men tested, 31 voluntarily withdrew from the program, 10 of them immediately

Table 4.1

Validity Coefficients and Related Data for Experimental Language Aptitude Tests Administered to Two Air Force Trial Course Groups

Entries are correlations or beta-weights with normalized academic grade criterion.

Test	Tested Feb. 1954 (N=80) *r*	Tested June 1954 (N=88) *r*	Selected Multiple *R*'s with Beta Coefficients		
			Feb. (a)	June (a)	June (b)
Artificial Language Part I	.33	.45			
Artificial Language Part II	.46	.52	.33	.23	.27
Artificial Language Part III	.45	.46			
Turse Spelling	.40	.47			
Turse Phonetic Association	.53	.62	.35	.16	
Spelling Clues	—	.53		-.04	
Turse Word Discrimination	.30	—			
Letter—Star	.24	—			
Word Squares	.19	—			
Disarranged Letters	.33	.48			
Rhyming	.32	.51			
Phrase Completion	.24	—			
Cooperative Vocabulary Form Z	.36	.42			
Artificial Language Numbers	.45	—			
Number Learning	—	.53		.27	.29
Words in Sentences	.51	.52	.23	.06	
Phonetic Discrimination P-120-A	.22	—			
Phonetic Discrimination P-123-A	—	.27			
Disarranged Words	.45	—			
Paired Associates	.50	.55	.23	.14	.19
Word Elements	.42	—			
Anagrams	.18	—			
Picture Naming	.06	—			
Verbal Enumeration: Number Attempted	.00	—			
Verbal Enumeration: Number Wrong (Refl.)	.21	—			
Same—Opposites: Number Attempted	.24	—			
Same—Opposites: Number Wrong (Refl.)	.26	—			
Devanagari Script	—	.37		-.05	
Perdašeb (Total)	—	.56		.14	
Phonetic Script (Total)	—	.68		.27	.40
r Required at 5% level	.22	.21			
r Required at 1% level	.29	.28			
r Required at 0.1% level	.37	.36			
Multiple *R* for the variables which have beta's listed in the column above			.75	.84	.82

after the first day; 47 were failed, and 33 were finally selected for study at IFEL. The 31 cases of voluntary withdrawal were excluded from further analysis (there was no significant contingency between voluntary withdrawal and test scores), and validity coefficients were based on the remaining 80 cases.

Two criterion measurements were employed: the academic grades (transformed to approximate normality) assigned by the instructors to both cases selected and not selected for further study, and selection or non-selection. Selection or non-selection was highly correlated with the academic grades, but not perfectly, since those responsible for selection paid some attention to judged character and temperament. As shown in Table 4.1, a large number of test variables showed highly significant correlations with the criterion measurements; at the same time, there were a number of tests which manifestly did not have statistically significant validities. There were no striking or systematic differences between validities for the two criteria. When the academic grade criterion was employed, it was found possible to obtain a multiple correlation of .77, shrunken by Wherry's formula, from just six tests by a test selection computation. Four tests—Test 2 (Artificial Language Learning, Part II), Test 5 (Turse Phonetic Association), Test 14 (Words in Sentences) and Test 17 (Paired Associates)—yielded a multiple R of .75. An interesting aspect of the results was that the beta-weights were of the same general order of magnitude: .33, .35, .23 and .23, respectively. Not only did this first validation run produce startingly high validity coefficients, it also yielded a set of tests with high validities and relatively low intercorrelations, a situation idealized by the textbooks but seldom found in practice. The validities were so high that under the selection conditions which actually obtained, the trial course results and the prediction test results agreed on the classification of 66 out of 80 cases, or 82.5%.

These results obviously begged for replication and cross-validation. A battery of tests was administered to another group under the same conditions in June 1954; 103 men were tested, of whom 15 voluntarily withdrew before completion of the trial course, 44 were failed, and 34 were selected. This second group was highly comparable to the first in test scores, criterion academic grades, and percentage selected. The battery of tests administered to this group contained a few new ones which had been constructed in the light of previous results, and some of the tests previously found invalid were now omitted. The results are to be found in Table 4.1. Using the regression equation

developed on the basis of the best combinations of the tests in the February results, a correlation of .77 was obtained between predicted scores and actual (normalized) academic grades, as compared with a multiple R of .75 obtained for the original sample. If anything, this was negative shrinkage. When certain of the newer tests were utilized in the prediction formula, a correlation of .84 was achieved. In comparison with validity coefficients commonly obtained for aptitude tests and in view of the possibility of some unreliability in the criterion, these figures represented an unusual degree of success in the prediction of learning ability.

It should be noted, however, that the criterion itself had many elements of a test situation; in fact, the trial-course was conceived of by the instructors as a week-long test and men were almost mercilessly dropped at the first real sign of weakness. Furthermore, the tasks used in the training program were highly similar to some of the tasks used in the test itself. For example, one of the first things the students had to learn was to recognize the correspondences between a series of Chinese phonemes, including the famous "tones," and their representations in the phonemic transcription used in the course, e.g., to know how to pronounce syllables marked thus: *mā, má, mǎ, mà.* This was very similar to the task posed in the Phonetic Script Test which had been introduced as a part of the aptitude test battery. Another task which had to be mastered was that of constructing simple Chinese sentences conforming to one of several "sentence-types" with characteristic placement of subject, verb, object, etc. This required that the individual be able to recognize subjects, verbs, objects, etc., and this was tested quite directly by the Words in Sentences task. Similar statements could be made for many of the other tests which yielded high validity coefficients.

It was therefore of interest to investigate the test validities with reference to a less immediate criterion—grades received by the selected cases at IFEL. Grades after five weeks of training were available for 65 cases from the two groups. These grades could be predicted with a correlation of .34 from a regression equation based on the tests, and .54 from the academic grades which had been assigned at the end of the trial course. On the basis of the high degree of selection which had taken place, it may be estimated that the test battery would probably have had a validity of more than .52 if all persons in the trial course had been allowed to study at IFEL. This figure, though not at all as high as the validity obtained for the prediction of trial course grades,

would still represent a highly useful degree of prediction. The tests do not appear to be quite as valid as the trial course itself, but the former have many economic advantages in comparison to the relatively expensive trial course.

Further data were compiled by Air Force psychologists on the subsequent use of the aptitude tests in the prediction of success in trial courses. As will be explained below, on the basis of the first year's studies of test validity in a number of situations, five tests were selected to compose a semi-final battery; tentatively, the battery was known as the Psi-Lambda Foreign Language Aptitude Battery (Psi-Lambda being an abbreviation of psycholinguistic). This battery was made available to the Air Force for further predictive studies. One of these studies concerned the prediction of success in a trial course in Russian. A multiple R of .64 was obtained for four of the tests (Phonetic Script I, Words in Sentences, Spelling Clues, and Paired Associates) with four classes comprising 95 cases; cross-validation of the resulting weighted composite on the next six classes, comprising 151 cases, yielded a correlation of .70 (Harding, 1956a).

Results were not quite so good in the Air Force studies of the trial course in Chinese (actually, this was precisely the same course as the one that had been used in the previous studies at IFEL, but the Air Force studies concerned groups subsequently entering training with possibly different selective biases); the multiple correlation on 135 airmen was only .45 (Harding, 1956b). The composite which had been developed from the Russian sample correlated only .39 with the criterion—which in this case was simply acceptance or non-acceptance into the longer course. Harding suggested that possibly a different combination of abilities was required to study Chinese.

A further experiment conducted by Harding (Harding, 1958; Harding & McWilliams, 1957) compared the use of a four-week trial course with use of the aptitude tests in predicting final grades of students in a six-months' intensive course in Russian. The validity coefficients obtained for a language aptitude composite score, .44 and .42 ($N=42$), were comparable to those obtained for trial course grades, .39 and .53 ($N=62$ and 52), respectively for two samples. The estimated validities which the tests would have had if all examinees had been allowed to go into training were .72 and .64 for two samples. Harding (1958, p. 122) further concluded that aptitude tests are more efficient than trial courses because applicants can be more readily screened by this means. He writes that, "This finding is at variance with a commonly held opinion that a trial performance in

a language is the best predictor of subsequent performance." This finding is also somewhat at variance with the IFEL results cited above, showing that trial course grades predict later grades (after five weeks) somewhat better than aptitude tests. A conservative plan for the selection of language trainees would utilize language aptitude tests to screen the input into a trial course; under these conditions, very few members of the trial course would be withdrawn for incompetence and hence the trial course would represent very little wasted effort. This, in fact, is the plan now being used by the Air Force.

Further validity coefficients as such will not be recited here. (The almost endless replication necessary in prediction studies makes aptitude testing in some ways one of the less interesting fields of applied psychology.) The remainder of this section will be used to introduce evidence on collateral questions such as the non-specificity of language aptitude to type of language, the nature of language aptitude, the role of background variables such as age and sex, etc.

Two series of tests were conducted at ALS and at the Foreign Service Institute. Both these institutions offer intensive training in a variety of languages, in courses of up to 12 months in duration. Data from ALS bear on *the question of the non-specificity of language aptitude to type of language*, since enough cases were available to warrant separate validity computations for several groups of languages, but not for individual languages. The languages involved could have been grouped in several ways, but it was decided to group them in a cross-classification of language family and use of Roman characters in the writing system since it could be argued that validity might depend on these factors. Table 4.2 shows a series of validity coefficients against final grades, oral and written work being weighted equally, for five subtests of the Psi-Lambda Foreign Language Aptitude Battery and for a weighted composite which had been developed on the basis of results from the Chinese trial course. It may be seen that while the validity coefficients in two successive samples vary somewhat, there is little evidence of a consistent pattern in these variables. For example, while the validity of the test is at its lowest in predicting success in the "character languages" (Japanese, Chinese, Korean) in one sample, it is not as low in the second sample. These data support the hypothesis of the non-specificity of language aptitude, a hypothesis which is supported by many other tabulations of research data which have been made. This is to say that high, as well as low, validity has been recorded for many different kinds of languages.

Table 4.2

Validities of Part and Total Scores for the
Psi-Lambda Foreign Language Aptitude Battery

Two Groups Tested at ALS by Language Group, Final Grades Used as Criterion
Group I, N=211; Group II, N=374

Languages	Group	N	Psi-Lambda Test Parts I Number Learning	II Phonetic Script	III Spelling Clues	IV Words in Sentences	V Paired Associates	Total
Indo-European, A*	I	83	.47	.36	.41	.52	.36	.58
	II	163	.33	.51	.38	.44	.41	.54
Indo-European, B*	I	47	.54	.44	.28	.46	.46	.57
	II	92	.32	.29	.18	.47	.18	.35
Non-Indo-European, B*	I	77	.15	.29	.14	.28	.17	.27
	II	104	.39	.42	.26	.38	.33	.45
All languages	I	211	.34	.32	.28	.41	.30	.45
	II	374	.36	.42	.30	.43	.33	.49

*Indo-European, A are Indo-European languages using the Roman alphabet: Czech, French, German, Polish, Romanian; Indo-European, B use other alphabets: Bulgarian, Greek, Russian. Non-Indo-European, B are non-Indo-European languages not using the Roman alphabet: Chinese, Japanese, and Korean. The data given for "all languages" include a few cases studying Hungarian, a non-Indo-European language using the Roman alphabet.

Table 4.3

Comparative Validities of Part Scores and Total Scores for the
Psi-Lambda Foreign Language Test, Against Grades in Oral and Written Work,
Together with Associated Inter-Correlation Data

Two Groups Tested at ALS, Group I (N=211), Group II (N=374)

Group		Grade Intercorrelations: Oral 3rd week	Oral 7th week	Written 3rd week	Written 7th week	Psi-Lambda Test Parts* I	II	III	IV	V	Composite Score
Oral, Third week	I	1.00	.85	.68	.67	.23	.41	.23	.33	.26	.40
	II	1.00	.83	.71	.68	.27	.43	.35	.34	.27	.44
Oral, Seventh week	I	.85	1.00	.72	.77	.32	.43	.31	.41	.29	.49
	II	.83	1.00	.71	.76	.33	.46	.39	.37	.25	.48
Written, Third week	I	.68	.72	1.00	.81	.34	.34	.23	.31	.26	.40
	II	.71	.71	1.00	.78	.33	.47	.36	.42	.30	.51
Written, Seventh week	I	.67	.77	.81	1.00	.40	.35	.27	.42	.28	.48
	II	.68	.76	.78	1.00	.34	.46	.41	.42	.27	.52
Final Grades	I	.45	.59	.46	.48	.34	.32	.28	.41	.30	.47
	II	.59	.62	.52	.54	.36	.42	.30	.43	.33	.51

*Test names are shown in Table 4.2

The data from ALS also bear on *the question of the differential predictability of oral and written work.* However, it should first be pointed out that grades in oral and written work at the school are highly correlated; the correlation between oral and written work grades at the seventh week of training was .82 in one sample of 251 cases that had been tested for aptitude at the outset of training. Table 4.3 shows the comparative validities of five tests and the weighted composite score. The differences between the correlations are relatively small. Possibly the difference for Phonetic Script is of theoretical as well as statistical significance, because it is a test involving sounds and the learning of symbols to represent sounds; this is probably an ability which is required to a greater extent in oral aspects of language learning than in written aspects. In interpreting this result, however, one should consider the fact that the high correlation between oral and written grades is probably a function of the way languages are taught at ALS and in similar programs. Both the oral and the written aspects are taught simultaneously or at least in close succession; the common element is therefore the language itself, its structure and its lexicon. Students must make approximately equal progress in oral and written work if they are to maintain their standings. Only in courses in which the reading objective or the speech objective is stressed nearly to the exclusion of the other objectives will there be a possibility of a low correlation between the separate kinds of evaluation; at the same time it may be expected that the aptitude test will best predict whichever criterion, oral or written, is stressed in the instruction. (Data bearing on this can be adduced from civilian academic settings, but will not be introduced here.)

Two sets of data are available from experimental testings at the Foreign Service Institute (FSI), Department of State. One interesting result pertained to *the nature of the criteria.* Thus far, little has been said about the criterion variables used in the studies reported; chiefly they have been grades assigned by instructors on a combination of subjective and objective bases. No efforts have been made to ascertain their reliability, but reliability must be quite high in view of the high validity coefficients they often engender. In the FSI testing, end-of-course criterion ratings were obtained on two bases: (*a*) "actual overall accomplishment," and (*b*) "estimated" aptitude and facility in language learning. The instructors were asked to give ratings on a hypothetical scale with 50 as the mean and 10 as the standard deviation; actually, they tended to spread their ratings somewhat more widely than this

around a mean in the neighborhood of 54 to 59. The courses were in 12 different languages and were six months in length.

In the first testing, which took place before the condensed battery took shape and thus involved a considerable number of tests, further evidence on the relative validity of the tests was obtained, as shown in Table 4.4. The Foreign Language Aptitude Index, computed from

Table 4.4

Validity Coefficients for Language Aptitude Tests for 68 Persons at the FSI

Test	Criteria (End-of-course Ratings by Instructors)	
	Actual Overall Accomplishment	Estimated Ability
Number Learning	.48	.51
Phonetic Script, first 30 items	.68	.68
Phonetic Script, last 30 items	.60	.61
Phonetic Script, total	.67	.67
Words in Sentences (Number attempted in 12½ minutes)	.37	.41
Words in Sentences (Number right in 20 minutes)	.62	.67
Artificial Language Learning (Total)	.60	.64
Spelling Clues	.58	.62
Oriental Script	.26	.31
Disarranged Letters	.50	.54
Paired Associates	.50	.52
Devanagari Script	.64	.62
Foreign Language Aptitude Index (weighted composite of Paired Associates, Number Learning, and Phonetic Script Total)	.69	.70

an integral-weight combination of five test scores, and developed in one of the Air Force trial course samples, had a correlation of .69 with the "actual accomplishment" criterion and .70 with the "estimated ability" criterion. While this difference is obviously not significant, it suggests that instructors in intensive language courses are not only able to make accurate judgments of accomplishment but also make allowances for factors which may attenuate the criterion, such as motivation, vicissitudes of personal life, etc. More use could be made of a criterion in which instructors are asked to estimate the ability demonstrated in a learning situation.

A question might be raised about the role of motivation to do well on the test. In the Air Force testings described earlier, it was distinctly the case that the examinees were under the impression that the test batteries would have a role in determining their selection for the course. This condition probably did not hold for the first FSI group, since most of the persons tested had already been selected for training. An attempt was made, therefore, to study *the effect of test anxiety* by administering, after the completion of the test battery, the Test Anxiety questionnaire developed by Sarason and Mandler (1952) who generally have found low but significant negative correlations between anxiety and performance on intelligence tests. Questions 3 to 13 on Sarason and Mandler's questionnaire ask about various feelings and attitudes which may be taken to reveal test anxiety; they concern mainly intelligence tests. A further question (number 14) was added, asking specifically how anxious the examinee felt during the aptitude tests he had just taken. Table 4.5 shows the correlations between these questionnaire results, the lan-

Table 4.5

Correlations between Test Anxiety Questions, Language Aptitude Index, and Criterion Scores

$N = 68$ FSI Language Trainees

	Mean	σ	1	2	3	4	5
1. Total Test Anxiety Score items 3-13 (possible range, 11 to 99)	31.6	15.0		.35	–.14	–.20	–.22
2. Question 14 (anxiety about this test, range 1 to 9)	2.3	1.7	.35		–.19	–.10	–.05
3. Language Aptitude Index (T-score scale)	60.3	10.3	–.14	–.19		.69	.70
4. Criterion: Actual Accomplishment	59.0	14.8	–.20	–.10	.69		.96
5. Criterion: Estimated Ability	58.0	15.0	–.22	–.05	.70	.96	

guage aptitude index, and the two criterion variables. General test anxiety had low negative correlations with test scores and with the criterion. The multiple correlation of total test anxiety and the language aptitude index as predictors of the "estimated ability" criterion is .71, a non-significant increase over the language aptitude index zero-order correlation of .70. The question asking whether the examinee felt any anxiety in taking the language aptitude test showed little relation to either test scores or criteria; the mean score, 2.3 on a scale from 1 to

Table 4.6

Intercorrelations, Means, and Standard Deviations for Psi-Lambda Foreign Language Aptitude Battery Subtests, Foreign Language Aptitude Index, Age, Prognostic Interview Ratings, and Criterion Ratings after Language Training

$N = 83$ Trainees at the FSI

	Predictor Variables								Criteria	
	1	2	3	4	5	6	7	8	9	10
1. Number Learning		.58	.57	.59	.53	.79	–.27	.40	.60	.60
2. Phonetic Script	.58		.69	.62	.38	.78	–.23	.52	.69	.75
3. Spelling Clues	.57	.69		.62	.46	.82	–.28	.36	.57	.64
4. Words in Sentences	.59	.62	.62		.51	.87	–.22	.29	.68	.72
5. Paired Associates	.53	.38	.46	.51		.74	–.29	.14	.46	.52
6. Foreign Language Aptitude Index	.79	.78	.82	.86	.74		–.32	.40	.74	.80
7. Age	–.27	–.23	–.28	–.22	–.29	–.32		–.00	–.21	–.18
8. Prognostic Interview Rating (reflected)	.40	.52	.36	.29	.14	.40	–.00		.55	.54
9. Criterion: Accomplishment	.60	.69	.57	.68	.46	.74	–.21	.55		.91
10. Criterion: Estimated Ability	.60	.75	.64	.72	.52	.80	–.18	.54	.91	
Mean	53.96	56.58	53.34	52.12	51.15	54.37	34.23	2.10*	56.47	54.00
σ	10.94	9.32	10.75	11.09	11.16	12.00	4.99	1.21	17.27	17.22

* For unreflected interview ratings, mean is 2.90. Original ratings were on basis of 1 = highest, 5 = lowest.

9, showed that there was very little test anxiety in any case. Anxiety, as measured, apparently did not affect the validity of the test, and it can hardly be said that the test is a measure of anxiety, despite the fact that the language aptitude score was correlated to the extent of −.19, not significant at the 5% level, with the self-appraisal of anxiety on the test.

A second group of 83 trainees, 77 men and six women, at the FSI was tested at the outset of training. No statements were made to the group with regard to whether the tests would be used in further selection. In advance of the testing, each examinee was also given an individual 15-minute "diagnostic interview" by the chairman of the language department in which he was to study. A number of questions on background, previous experience, and motivation were asked, but the ratings of probable language aptitude, on a five-point scale, were based chiefly on responses to informal mimicry tests in which candidates had to imitate short spoken phrases in a foreign language as accurately as possible. At the end of the six- or eight-month intensive courses into which the examinees were placed, the instructors made criterion ratings similar to those made for the earlier group. The results (means, standard deviations, and correlations of all variables, including age of subjects) are shown in Table 4.6. The scores on the five subtests of the Psi-Lambda Foreign Language Aptitude Battery are in terms of a *T*-score scale which had been developed from a preliminary standardization sample of 912 cases from a variety of sources.

For a multiple regression analysis of the five subtests as predictors of the two criterion variables, the zero-order correlations, beta-weights, and multiple correlations are as follows:

	Criterion 1 (*Accomplishment*)		*Criterion 2* (*Estimated Ability*)	
Test	*r*	*β*	*r*	*β*
Number Learning	.60	.16	.60	.07
Phonetic Script	.69	.38	.75	.43
Spelling Clues	.57	−.01	.64	.04
Words in Sentences	.68	.33	.72	.32
Paired Associates	.46	.07	.52	.13
Multiple *R*	.78		.83	

The validity coefficients obtained in this sample are among the highest that have been attained with any sample. Yet, the criterion was performance on a long, intensive course rather than simply a short trial course. It should be commented that the group was quite heterogeneous,

in that it included prospective foreign service officers and civilian employees of various government agencies as well as Air Force enlisted men. Unfortunately, it was impractical to administer a measure of verbal intelligence to this group, but there was probably a considerable range of intelligence represented.

The beta-weights tended to confirm the previous findings that the various subtests had important independent contributions to validity. Some of the beta-weights, of course, were close to zero, but experience has indicated that these weights are subject to considerable sampling fluctuation and may also be responsive to differences in types of courses and criteria used. It has seemed safest to leave all five subtests in the final battery in order to attain maximal overall validity in the long run and to provide prospective users with the possibility of performing their own studies of the differential validity of the subtests.

Another question answerable from the data presented in Table 4.6 is that of the relative usefulness of the language aptitude test, the prognostic interview ratings, and age. The zero-order correlations and beta-weights bearing on this question are as follows:

	Criterion 1 (*Accomplishment*)		*Criterion 2* (*Estimated Ability*)	
Predictor	*r*	*β*	*r*	*β*
Foreign Language Aptitude Index	.74	.62	.80	.71
Interview Rating	.55	.30	.54	.26
Age	−.21	−.01	−.18	.05
Multiple *R*	.79		.84	

It is evident that while the interview rating contributes a useful amount, significant at the .1% level in each case, over and above the language aptitude index, its validity comes nowhere near surpassing that of the test score. This is true despite the fact that the interview rating may contain a spurious overlap with the criterion ratings in that many of the interview ratings were made by the same individuals who later awarded the criterion ratings. In many situations, of course, use of a prognostic interview would be impractical, either because of the large numbers of cases to be handled or because of the unavailability of qualified, linguistically-trained interviewers at the site of testing. Nevertheless, the results do indicate that interviewing which includes a mimicry test may be of some help in selecting language trainees. One interesting facet of the results is that the prognostic interview rating correlates most highly with the Phonetic Script Test, suggesting that there is

much in common between what is measured by this test and the informal mimicry exercises given in the prognostic interview. This matter will be discussed in the section on the nature of language aptitude.

Among the samples studied, this group is unusual in that its mean age is 34.23, σ= 4.99. Age shows a slightly negative linear correlation with success in language learning, but since the age variable does not contribute to prediction over and above the aptitude test, it may be assumed that the aptitude test measures whatever in the age variable is relevant for language training success. These results tend to deny the popular notion that older individuals cannot learn foreign languages readily.

There were two notable examples of situations where the language aptitude tests showed very poor or even negligible validity. One of these was at the National Security Agency (NSA) of the Department of the Army, but here the criterion was very poorly defined, or perhaps it was irrelevant to what the language aptitude tests were supposed to measure. It was pointed out that the criterion grades at NSA represented the extent to which the students were able to learn to use foreign language skills in cryptanalysis and related matters. Many of these already had prior training in foreign languages. In any case, 62 persons were given a large battery of tests at the outset of certain training courses, and at the end of the courses, typically six months in duration, test scores were correlated with course grades. Scores on several other tests were available for 37 of these cases. A small portion of the results is given in Table 4.7. Of chief interest is the fact that none of the validity coefficients approaches significance.

The correlations between selected language aptitude tests and other tests merit attention. It may be noted that the Words in Sentences Test appears to have a consistently high correlation with intelligence tests. Its correlation, .87, with the Language Inference part of the Iowa Foreign Language Aptitude Test is quite high; this is a part in which it is required that the examinee infer, from context and from knowledge of cognates, the meanings of Esperanto words like *luno*, *urbo*. The possibility that these two tests are strongly affected by previous language training experiences should be investigated. Other tests, such as the Phonetic Script Test and the Paired Associates Tests, are not highly correlated with subtests in the Iowa test, although there are a few moderate correlations; these tests, at least, seem to be rather dissimilar to previous language aptitude tests.

Table 4.7

Intercorrelations Between Language Aptitude Tests, Intellectual Ability Tests, and Course Grades

$N = 37$ cases at the NSA

Test	Coop. Vocab.	Words in Sentences	Phonetic Script	Disarranged Letters	Paired Assoc.	Course Grade (Criterion)	Mean	σ
ACE Quantitative	.22	.47	.16	.51	.21	.14	50.97	9.55
ACE Linguistic	.56	.58	.35	.27	.37	.25	99.92	9.74
ACE Total	.46	.62	.30	.46	.34	.22	150.97	16.38
Iowa FL Apt.: Lang. Inference	.13	.87	.30	–.11	.18	.08	27.00	4.42
Iowa FL Apt.: Lang. Construction	–.09	.51	.26	.50	.41	.20	80.24	12.55
Iowa FL Apt.: Grammar	–.20	.56	.13	.11	.41	.13	67.73	8.78
Iowa FL Apt.: Total	–.11	.51	.25	.07	.43	.17	174.70	22.36
Coop. Read. Comp.: Vocabulary	.74	.30	.27	.33	–.06	.15	74.95	5.64
Coop. Read. Comp.: Speed	.57	.52	.15	.46	.28	.07	77.65	13.43
Coop. Read. Comp.: Level	.53	.42	.14	.44	.15	.15	72.27	10.62
Course Grade (N=62)	.12	.12	.12	.03	.14	1.00	5.48	1.24
Mean	78.38	46.46	52.57	23.97	19.76	5.86		
σ	6.18	7.03	7.19	6.32	4.62	1.32		

The other situation in which the attempt to predict language learning grades was largely unsuccessful was at a university specializing in the intensive teaching of Russian to U. S. Air Force personnel. Two classes were given the same experimental battery that had also been given to the Air Force "trial course" group. Class II had just begun Russian language training, while Class I was being tested at the beginning of its second term of Russian language study. The examinees were all enlisted Air Force personnel who had been preselected like the members of the Chinese trial course group, and they had also been screened by a similar trial course, but in Russian. The criterion scores consisted of oral and written grades for the first five marking periods from October to March for Class I, and for the first two marking periods for Class II. Table 4.8 shows selected data from this study; data are given only for tests which are identical or similar to tests which were later chosen for the final battery. Actually, the only test which yielded validity coefficients significantly different from zero at the 1% level was the Turse Phonetic Association Test, which seems to measure much the same trait or traits as Spelling Clues and Phonetic Script in the final form of the battery, and even these results were not consistent for the two classes.

Table 4.8

Selected Validity Coefficients for Air Force Russian Language Trainees, with Intercorrelations of Oral and Written Criteria in the Several Marking Periods

Variable	Class 1 (N=46) Average Grades				Class 2 (N=30) Average Grades	
	Oral		Written		Oral	Written
	1, 2	3, 4, 5	1, 2	3, 4, 5	1, 2	1, 2
Tests:						
Turse Phonetic Association (cf. Spelling Clues)	.43*	.22	.13	.40*	.02	–.13
Artific. Lang. Numbers (cf. Number Learning)	.02	.05	–.24	–.17	–.02	–.16
Words in Sentences	.12	.06	.03	.01	.05	–.03
Paired Associates	.15	.07	.10	–.01	–.08	–.22
Grades:						
Oral, Period 1, 2	1.00	.74	.68	.69	1.00	.09
Oral, Period 3, 4, 5	.74	1.00	.60	.70	—	—
Written, Period 1, 2	.68	.60	1.00	.68	.09	1.00
Written, Period 3, 4, 5	.69	.70	.68	1.00	—	—

*Significant at the 1% level.

A number of hypotheses to account for the low validity coefficients may be considered, including the restriction in range due to the prior selection by a trial course, poor motivation due to the groups' already having been selected, and inadequacy of the criteria. Analysis shows that restriction in range can account for only a slight amount of the decrement in validity. No data are at hand concerning the effects of poor motivation. The inadequacy of the criterion is probably the most reasonable explanation; the negligible correlation between oral and written grades in Class II leads one to cast considerable doubt on the criterion at least for that class. Associated with the criterion should be considered such matters as the quality of the teaching, the quality of the text materials, and the reliability of grading.

The fact that the tests do not always predict a given set of criterion ratings does not mean necessarily that the tests are invalid. The data collected tend to show that the tests assembled in the battery are, generally speaking, highly valid, and it can be inferred that they measure some complex of traits which make for success in language learning. Individuals who score low on the tests will sometimes do well in a classroom, but it is hypothesized that such an individual will have to work harder than a high-scoring student, or that such a student must have a more than ordinarily patient teacher. Later in this chapter a theoretical discussion will be presented of the conditions under which aptitude test scores show high correlations with criterion measures.

Use of the Tests in the Diagnosis of Learning Difficulties

If the tests do indeed provide measures of several somewhat independent abilities involved in language learning, it follows that the subtest scores ought to contain information which would be useful for the diagnosis and prediction of specific learning difficulties. An opportunity to investigate this in the context of an intensive language training course was presented by the so-called Five-University Summer Program in Middle Eastern Languages. In 1958 and 1959, groups of students in this program were tested at the beginning of the summer, and the scores were compared with the results of graphic rating scales filled out by instructors at the end of the eight-week intensive courses in Arabic, Turkish, Persian, or Modern Hebrew. The rating forms called for assessments of highly specific aspects of language learning behavior. In the first summer, when they were handed out near the end of the course, instructors complained that they had not been observing individual students closely enough to make accurate ratings in every case; in the second sum-

mer, therefore, they were distributed to instructors well in advance of the end of the course.

The correlations (one from each year) of each rating scale with each part of the Modern Language Aptitude Test (MLAT), the commercial version of the Psi-Lambda Test (Carroll & Sapon, 1958), are shown in Table 4.9. Correlations are also shown for three questions which the students answered on a questionnaire filled out at the outset of the course. Despite the small number of cases, the correlations for the two years show a reasonably consistent pattern; the correlations for 1959 tend to be somewhat higher. From the standpoint of overall validity against final course grades, the results presented here are promising. The correlations of total test with final grade are .40 and .58 respectively, for the two summers. Validities against mean diagnostic rating are even higher, .55 and .69, respectively; this result is reminiscent of that obtained at the FSI wherein ratings of "estimated ability" were predicted slightly better than ratings of actual performance.

The tests also appear to have some degree of diagnostic significance. Each diagnostic rating is related to a particular test or combination of tests. Thus, ability to hear phonemic distinctions seems to be related more consistently to the Phonetic Script Test than to any other. Ability to produce phonemes accurately and to mimic basic sentences seems most closely related to Spelling Clues. Memorizing vocabulary is most closely associated with the Number Learning and the Paired Associates Tests which are, in fact, memory tests. It had been thought that ratings of ability to understand grammar and to speak grammatically would have closest relationships with the Words in Sentences Test, and while there is indeed a consistent relationship, they are also predicted reasonably well by the Phonetic Script Test, or even by the Number Learning Test. Some ratings are about equally well predicted by all subtests. Of course, it is hard to know whether to take the ratings themselves at face value, since there is no way of guaranteeing that an instructor would be able to separate the several aspects of behavior even conceptually. Nevertheless, the results shown here suggest that a careful diagnostic study of test scores can be worthwile in predicting later learning difficulties.

The last three rows of the table show that: (*a*) whether or not a person likes foreign language study is not related significantly either to aptitude or to achievement; (*b*) the subject's statement of whether he has found foreign languages easy or hard is moderately well related to certain aspects of aptitude (most consistently to the Number Learning, Phonetic Script, and Words in Sentences Tests) and also to

Table 4.9

Correlations Pertinent to the Use of Language Aptitude Tests for Differential Diagnosis; Results from Testing in the Five-University Summer Program in Mideast Languages*

Variables	Number Learning		Phonetic Script		Spelling Clues		Words in Sentences		Paired Associates		MLAT Total		Total Course Grade	
Total course grade	.41	.58	.26	.49	.31	.42	.42	.41	–.07	.59	.40	.58	1.00	1.00
Instructors' ratings on ability to:														
Hear phonemic distinctions	.40	.51	.56	.58	.32	.58	.51	.39	.25	.43	.59	.59	.64	.81
Produce phonemes accurately	.37	.49	.38	.59	.48	.52	.34	.35	.05	.45	.45	.56	.63	.72
Mimic basic sentences	.22	.30	.13	.49	.46	.41	.13	.21	.29	.39	.36	.41	.39	.66
Memorize vocabulary	.61	.64	.54	.58	.31	.43	.56	.51	.14	.75	.63	.68	.71	.78
Understand grammar	.47	.51	.52	.56	.21	.42	.50	.45	.03	.58	.50	.58	.75	.85
Speak grammatically	.57	.59	.55	.60	.37	.47	.62	.56	.15	.57	.67	.66	.67	.85
Comprehend spoken language	.47	.56	.48	.49	.43	.50	.43	.46	–.02	.48	.51	.60	.76	.75
Comprehend written language	.28	.67	.41	.55	.00	.38	.45	.52	.12	.68	.38	.65	.39	.88
Mean of above ratings	.48	.65	.46	.61	.39	.53	.50	.51	.07	.64	.55	.69	.82	.90
Student responses to questionnaire:														
Liking for foreign languages	.24	.19	.16	.26	–.01	.16	.10	.31	–.27	.34	.04	.29	.12	.16
Judged ease of foreign languages	.57	.42	.57	.35	–.05	.11	.63	.45	.03	.50	.52	.42	.41	.31
"Academic compulsiveness"	.24	–.06	.24	–.27	–.04	–.37	.14	–.37	.21	–.11	.21	–.29	–.17	–.05

*Two values in each cell are presented: the first is for $N = 32$ students in the 1958 program, the second is for $N = 30$ students in the 1959 program. Approximate significance levels: $r = .36$ for 5% level; $r = .48$ for 1% level.

achievement as measured by the final grade in the course; (*c*) there is relatively little significance in the question about "academic compulsiveness," i.e., self-classification into Type A or Type B students as described below.

Validity of Language Aptitude Tests at the Service Academies

Although this chapter is concerned chiefly with the prediction of success in intensive "full-time" language learning, it has been indicated that it is not irrelevant to consider results obtained at the two service academies, the U. S. Military Academy at West Point, and the U. S. Air Force Academy in Colorado. The fact is that, at least at West Point, the cadet spends somewhat more time on language learning in his first two years than is the case at many colleges. Further, the language instruction at both academies tends to emphasize speaking and understanding rather than reading and writing.

A large test battery was administered in 1954 to a total of 619 "plebes," cadets in their freshman year, approximately four months after foreign language training had begun. These cadets had been assigned to study the language of their choice as far as possible within the quota set up for each language: 114 in French, 110 in German, 54 in Portuguese, 107 in Russian, and 234 in Spanish. The tests were the same as those also administered to the Air Force Chinese trial course group and the Air Force Russian language training groups, except that because of time limitations only half of the tests could be given to any one cadet. In addition, a questionnaire was administered to yield data on background, previous foreign language contacts, and motivational factors. Further, two scores were available on the West Point Aptitude Test (WPAT) given prior to entrance, the Total Score and the Language Aptitude Score.

The criterion data consist of first-term grade averages, on a percentage scale, collected nearly concurrently with the testing, and also the language average for the complete two years of language study, on the West Point grading scale of 0.0 to 3.0. In addition, academic standings at the end of the first two years were collected for English and mathematics as well as the Cumulative Order of Merit, which reflects both academic and certain non-academic kinds of performances at West Point. The first-term language grades were adjusted in such a way as to eliminate the effect of prior training in the respective languages.

Table 4.10

Correlations of Experimental Language Aptitude Tests with the West Point Aptitude Test and with Grades in Foreign Languages and Other Subjects at the U. S. Military Academy

Test or Variable	N	West Point Aptitude Test: Lang. Score	West Point Aptitude Test: Total Score	Foreign Lang. Grades: First Term	Foreign Lang. Grades: Two-Yr. Aver.	Two-Yr. Averages in: Eng.	Two-Yr. Averages in: Math	Cum. Order of Merit (refl.)
Artificial Lg. Learning I	242	.16	.22	.17	.16	.22	.16	.23
Artificial Lg. Learning II	242	.17	.15	.14	.14	.20	.11	.18
Artificial Lg. Learning III	242	.10	.23	.14	.11	.19	.16	.20
Turse: Spelling	242	.07	.24	.24	.30	.41	.12	.22
Turse: Phonetic Association	242	.13	.30	.39	.40	.45	.13	.26
Turse: Word Discrimination	242	.18	.36	.26	.30	.54	.10	.24
Letter—Star Test	297	.06	.13	.17	.24	.22	.16	.24
Word Squares	297	.02	.01	.01	.02	.00	.10	.12
Disarranged Letters	297	.13	.22	.17	.26	.36	.10	.23
Rhyming	297	.16	.29	.26	.31	.41	.13	.26
Phrase Completion	297	.17	.22	.11	.13	.33	.04	.16
Coop. Vocabulary	297	.16	.35	.21	.22	.42	.08	.21
Artificial Language Numbers	277	.25	.31	.28	.30	.26	.25	.32
Words in Sentences	277	.26	.27	.18	.25	.35	.28	.36
Phonetic Discrimination P-120	277	.08	.06	.10	.07	.02	.02	.06
Disarranged Words	291	.31	.37	.27	.30	.40	.14	.27
Paired Associates	291	.24	.21	.20	.20	.25	.12	.19
Word Elements	291	.31	.34	.30	.30	.36	.26	.33
Anagrams	291	.17	.11	.14	.13	.21	.07	.12
Picture Naming	291	–.02	–.06	.12	.11	.12	.06	.08
Verbal Enumeration: No. attempted	291	.09	.10	.08	.07	.04	.04	.05
Verbal Enumeration: Wrong (Refl.)	291	.06	.03	.01	.10	.14	–.02	.06
Same—Opp.: No. attempted	291	.20	.15	.13	.20	.18	.17	.20
Same—Opp.: Wrong (Refl.)	291	.07	.04	.02	.05	.12	–.01	.03
WPAT: Language Score	611*	1.00	.60	.31	.33	.32	.24	.31
WPAT: Total	611*	.60	1.00	.32	.36	.46	.45	.49
Foreign Language Grade: 1st Term	611*	.31	.32	1.00	.82	.44	.41	.54
Foreign Language Grade: 2-yr. Aver.	611*	.33	.36	.82	1.00	.54	.47	.65
English Grade: 2-yr. Aver.	611*	.32	.46	.44	.54	1.00	.39	.59
Mathematics Grade: 2-yr. Aver.	611*	.24	.45	.41	.47	.39	1.00	.92
Cumulative Order of Merit (Refl.)	611*	.31	.49	.54	.65	.59	.92	1.00
Composite Language Apt. Score A**	—	.32	.40	.38	.42	.48	.29	.42
Composite Language Apt. Score B**	—	.59	.53	.42	.46	.51	.32	.45

*Correlations in these rows were computed by averaging (via Fisher's *z*-transformation) correlations obtained for four separate but overlapping subgroups.

**Composite A contains (with unit weights for standard scores) Artificial Language Numbers, Turse Phonetic Association, Words in Sentences, and Paired Associates. Composite B contains the preceding four tests plus the WPAT Language Score.

The results are presented in Table 4.10. The five language groups were pooled since analysis of data for separate language had not shown consistent differential patterns. The following observations may be made about the results:

1. None of the experimental series of tests correlates even moderately well with the Language Score of the West Point Aptitude Test; the highest correlation is .31 and though it is significant at the 1% level, it is not such as to suggest any substantial degree of overlap with that test. The correlations of the experimental tests with the Total Score of the WPAT are somewhat higher, but only for those tests with a considerable verbal component such as the Cooperative Vocabulary Test or the Word Discrimination subtest of the Turse Stenographic Aptitude Test.

2. The correlations of the experimental tests with West Point foreign language grades are generally low, but several tests have validities high enough to be of predictive value, namely, the Phonetic Association subtest of the Turse Test, the Rhyming Test, and the Artificial Language Numbers Test. The first of these tests had a validity coefficient of .40 which is higher than that of the WPAT Total Score, .36, against the two-year grade criterion. This test battery, of course, was the first battery to be constructed and reported on. It did not contain several tests, notably, the Phonetic Script Test, which later proved to have an important role in the final prediction battery. Nevertheless, if a composite is formed from the four tests most closely approximating the final battery (Turse Phonetic Association, Artificial Language Numbers, Words in Sentences, and Paired Associates), the correlations with first-term language grades and the two-year average will approximate .38 and .42 respectively. These correlations are increased to .42 and .46 if the WPAT Language Score is added to the composite. (These composites are formed by assigning unit weights to standard scores.)

3. Some of the tests are also fairly good predictors of English grades, particularly those involving a clear verbal component. There is only one test, Artificial Language Numbers, which tends to have a higher correlation with language grades than with English grades.

4. The language aptitude tests are distinctly poor predictors of mathematics grades, as they should be.

It is quite likely that there are extraneous factors at West Point which serve to depress the validities of the language aptitude tests. Because of the strictly regulated study time available to the student, success in a foreign language class may easily be affected by the amount of

study time which the student feels he has left over from other courses such as mathematics, which may loom larger to him as a factor determining his career at West Point and which thus demand more of his attention. It is possibly partly for this reason that the WPAT Total Score is a better predictor of language grades than the language score itself, because the Total Score predicts success in other courses besides foreign language courses.

Data on the validity of the Psi-Lambda Foreign Language Aptitude Battery for the Class of 1959 at the U. S. Air Force Academy are presented in Table 4.11.[4] The validity coefficients are similar in magnitude to those obtained at the U. S. Military Academy. The relatively small degree of correlation can probably not be accounted for in terms of restriction of range, since the Air Force sample shows a mean *T*-score close to the mean of a preliminary standardized group known to be quite heterogeneous.

Table 4.11

Validity Coefficients for Psi-Lambda Foreign Language Aptitude Battery (Total Weighted Score) at U. S. Air Force Academy

Language	*N*	*r* with First Semester Grades	*r* with End-of-Year Grades	Mean Psi-Lambda (*T*-score scale)
German	19	.53	.53	50.0
Russian	35	.30	.24	54.8
French	56	.23	.30	49.0
Spanish	91	.30	.32	47.5
All languages	201	.30	.34	49.2

A Model for Studying the Prediction of Success in Complex Learning Tasks

Up to the present point in this chapter, a very simple model for studying the prediction of success in foreign language training has been assumed. The model assumes that success is a direct function of measured aptitude, plus errors which may exist in the tests or in the criterion; thus, for a given individual, $c + e_c = f(a) + e_a$, where c is the true score for the criterion, a is the true score for aptitude, and e_a and e_c are symbols for errors in c and a. Aptitude, a, may be regarded as a composite of true scores for various aspects of aptitude. The validity coefficient, i.e., the

[4]Furnished through the courtesy of Lt. Col. William F. Long, Director of Admissions.

correlation between aptitude test scores and a criterion measurement, is regarded as an indication of the extent to which errors have been minimized.

Obviously this model is oversimplified, if not downright wrong. It might be approximately correct under certain conditions, as where students are equally well motivated to learn, and are given only as much opportunity to learn as is actually needed by the more apt students. The model also assumes that there is only one kind of aptitude which is relevant to task success no matter how the learning task is organized. Travers (1954) has suggested a model which involves both motivation and aptitude; he believes these factors should be combined by a multiplicative function. McBee and Duke (1960) have presented evidence for an additive function, however. It would seem desirable to develop a model which would take account of not only aptitude and motivation, but also the relevant instructional variables. The analysis to be presented here is stated in general terms so that it is not restricted to the case of foreign language learning.

Consider a complex learning task as composed of a series of subtasks, which may be learned with varying degrees of perfection depending upon a number of circumstances. The model attempts to suggest the nature of these circumstances and the manner in which they affect the degree of learning. The degree of learning is measured in terms of the amount of success achieved in the total learning task after a fixed amount of elapsed time, e.g., after a specified number of weeks or months of a course, or after a specified number of instructional hours. The independent variables needed in the model fall into two categories, variables associated with the conduct of instruction and variables associated with the individual.

Instructional Variables.

p_j = *adequacy of presentation of task j* (on a scale from 0 to 1). This is a measure of how clearly the task is presented and explained, and how appropriately it is placed in the sequence of graded tasks to be learned. Efforts to "program" instruction for teaching machines and the like are essentially efforts to maximize p_j for every task; good textbooks and good teachers also seek to maximize p_j's.

o_j = *the time allowed, "opportunity," for learning task j.* Opportunity is presumed equal for all individuals; at least $\Sigma_j o_j$ is constant for all individuals over a group of tasks collected into a course of instruction.

Individual Difference Variables.

g_i = *that characteristic, general intelligence or verbal intelligence, which determines the extent to which the individual will be able to understand directions and explanations or to infer such directions and explanations from the total content of the instruction even when they are lacking.* This variable is measured on the standard score scale (mean = 0, $\sigma = 1$) and is assumed to interact with p_j in such a way that $u_{ij} = f_1(g_i, p_j)$, that is, the individual's understanding of the task requirements or his cognitive orientation to the means of meeting them is a function of his general intelligence, g_i, and the adequacy with which the task is presented. The f_1 function is tentatively defined so that $u_{ij} = f_1[p_j/A(g_i)]$, where $A(g_i)$ is the normal curve area above the value of g_i, but also such that $u_{ij} = 1$ for all $p_j > A(g_i)$.

a_{ij} = *the time which would be needed by individual i to learn task j to a specified criterion of learning,* on the assumption that $u_{ij} = 1$ (i.e., that the task is presented well enough for him to understand the task in the light of his g_i). This variable represents "aptitude," and is assumed to be a relatively invariant characteristic of the individual, not subject to easy modification by learning. It is assumed further that there may be a net of functional relationships among the values of a_{ij} for a given individual over different values of j; these functional relationships may be represented by "factors" (as in factor analysis), and it is possible that a very small number of factors may account for the relationships to a given degree of precision. It should be noted that low values of a denote "high" aptitude, i.e., the individual needs little time for learning.

m_{ij} = *the maximum amount of time individual i would apply himself to the learning of task j.* This may in turn be a function of the amount of difficulty the individual perceives in the task, his "motivation," and other variables, but these subsidiary variables will not be directly represented in this model.

The attempt can now be made to postulate a functional relationship between a criterion of success in learning and the above variables. This relationship will still be very much oversimplified but it will help to study the separate effects of the several variables on the correlation between aptitude and the criterion. First, it will be useful to define several derived variables. As stated above, if $u_{ij} = 1$, it can be assumed that the time needed by an individual to learn a task is a_{ij}. But if $u_{ij} < 1$, he can still

learn the task, but this will require more time. If p_j takes any value from 0 to 1, the time actually needed by individual i to learn a criterion is

$$a'_{ij} = a_{ij}/u_{ij}.$$

Obviously, as p_j approaches zero, a'_{ij} approaches infinity.

In the course of learning a task, whether he ever reaches the criterion of mastery or not, the time the individual spends will be a function of a'_{ij}, m_{ij}, and o_j. In fact, the time he spends will be the smallest of these values, since it is assumed that the individual will stop work as soon as he either (*a*) learns the task to the specified criterion of mastery, (*b*) spends an amount of time denoted by m_{ij}, or (*c*) is precluded from completing his learning because of the expiration of time as denoted by o_j, whichever of these events occurs earliest. In this way, t_{ij} can be defined as the total amount of time spent by individual i in learning task j, or, $t_{ij} = Sm\ (a'_{ij}, m_{ij}, o_j)$, where the symbol Sm denotes the function "smallest of the values listed."

Finally, it can be assumed that the efficiency, c_{ij}, with which task j is learned by individual i is a direct function of the ratio of the time spent to the time needed; that is, $c_{ij} = t_{ij}/a'_{ij}$. The final criterion of success in a series of tasks ($j = 1, 2, \ldots, n$) could then be represented as equal to $\sum_j c_{ij}$.

It is now of interest to study the relations between c_{ij} and a_{ij} (and other variables) under varying conditions of p_j and o_j, the instructional variables, assuming normal distributions of g_i and a_{ij} and also assuming that g_i and a_{ij} are independent. For this purpose it is assumed that only one task is being studied, i.e., j is a constant, and will therefore not appear as a subscript in what follows. The method adopted is to construct synthetic, hypothetical data and then to compute various statistics from these data under varying assumed conditions of instruction and "motivation" as represented by the variable which has been designated m_{ij}. The number of possible combinations of initial conditions from which one might start is of course infinite; it has been necessary to choose several such combinations arbitrarily, but the results will nevertheless suggest the trends which may be expected under a variety of conditions. In constructing hypothetical data, a basic sample of 100 cases was established; it was assumed that a_i and m_i were distributed normally with a population mean = 5.0 and variance = 1.0, also that g_i was distributed normally and independently of a_i and m_i with a population mean = 0.0 and variance = 1.0; from these assumptions sample values were developed by random sampling techniques. The sample values of a, g, and m had means, variances, and intercorrelations

well within 95% confidence bands around their expected values. A program was written for the IBM 704 electronic data processing machine to compute means and σ's of c_i for various combinations of o and p. The results of two runs are displayed as Case 1 and Case 2 in Figure 4.1.

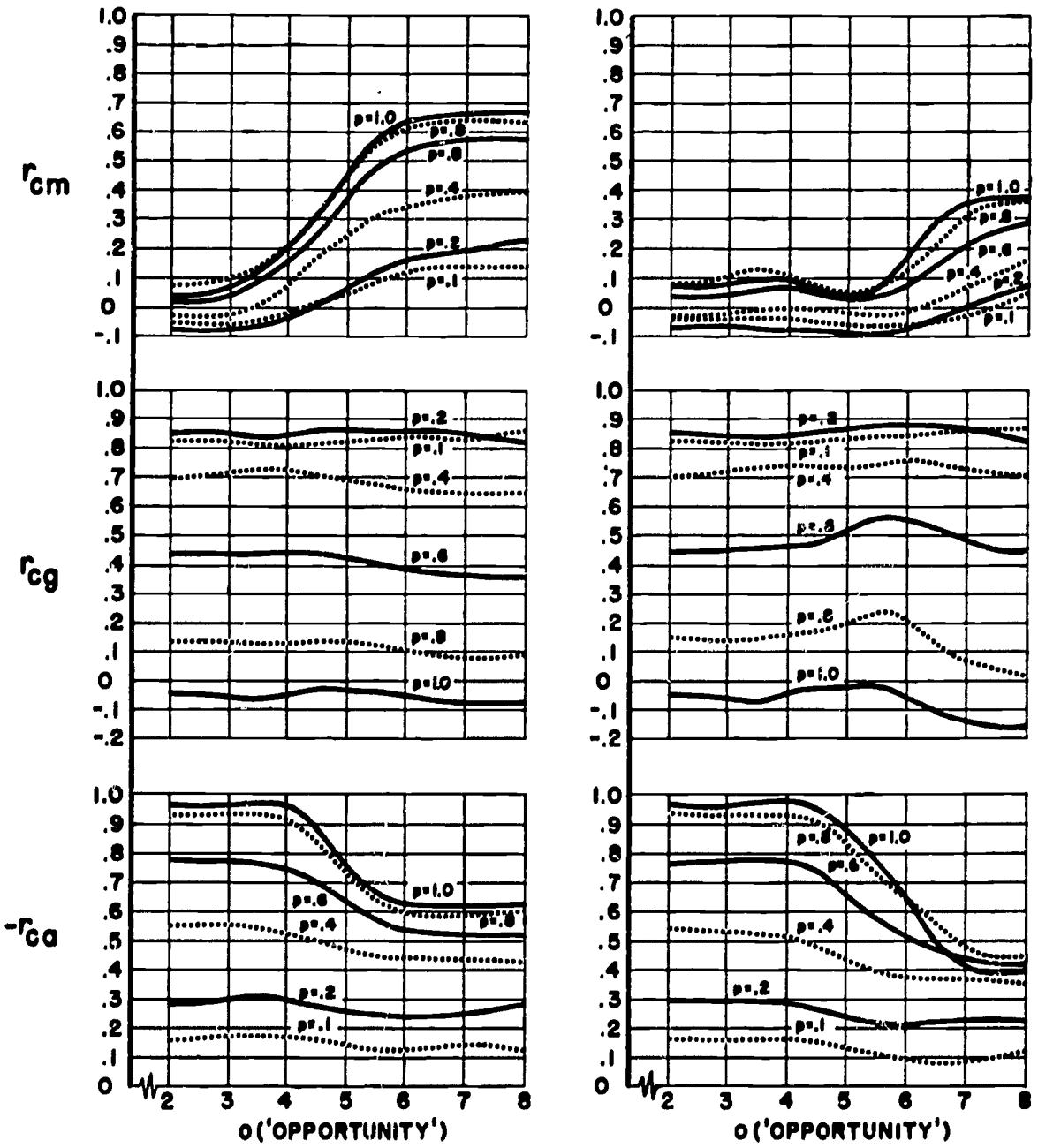

FIGURE 4.1 Computed Values of r_{ca}, r_{cm}, and r_{cg} for Various Values of o and p, for Two Levels of m (Hypothetical Data)

Case 1 is for a hypothetical group whose values of m_i, in the population, are distributed with the same mean and variance as their values of a_i, as indicated above. If m is regarded as a measure of motivation, it may be said that Case 1 refers to a group with relatively poor motivation; only about 50% of the group are willing to spend the amount of time in learning the task that they need. In computing the data for Case 2, however, a constant of two was added to each m_i. Thus, the data of Case 2 can be regarded as referring to a highly motivated group; for, given adequate opportunity and adequate instruction, nearly all individuals are posited to be willing to spend enough time to learn.

If o_j is varied, the amount of opportunity (measured in time) given for learning is varied; the less opportunity given, it may be expected that the less apt students will have less and less chance to catch up with more apt students. As o_j decreases, therefore, it would be expected that the correlation between aptitude ($-a_i$, that is, the amount of time needed, measured in reverse) and success increase, and this is seen to be true in both Case 1 and Case 2 in Figure 4.1. The relation between aptitude validity and opportunity to learn does not depend much on the average level of motivation in the range studied here. At the same time, the relation between success and degree of individual motivation is, as would be expected, generally higher in Case 1, that is, the group with a lower average level of motivation.

If p_j is varied, the quality of instructional presentation is varied; that is, the clarity with which the subject matter is presented and/or the appropriateness of the grading or ordering of material is varied. (p is conceived to be independent of o, e.g., material might be presented very clearly, but little time is given the student to assimilate it.) Since a low quality of instruction is in this model conceived to cause the learner to require more time to assimilate the material, depending upon his general intelligence, g_i, variations in p will affect the relations between success, on the one hand, and aptitude, a, general intelligence, g, and motivation, m, on the other. As p decreases from the optimum, aptitude has less relation to the criterion and general intelligence has much more. Likewise, the influence of individual differences in motivation is attenuated. (In the model, p is conceived to be independent of m, although in practice, it is possible that low p lowers student motivation.)

This model has been designed chiefly to illustrate the possible effects that instructional variables may have on the functioning of an aptitude test. Some of its assumptions are over-generous, e.g., the assumption that general intelligence is uncorrelated with aptitude and that it operates

only in orienting individuals to understand tasks. Nevertheless, it can be used to interpret some of the results presented in this chapter.

It will be recalled that the highest degree of relationship between aptitude and achievement was demonstrated for the two U. S. Air Force trial courses. It is assumed that in this course the value of o_j was relatively low, probably near the point where the relationship between c_{ij} and a_{ij} would be at a maximum. It is entirely credible that o_j was relatively low because of the extreme brevity of the course and the fact that the instructors were almost ruthless in eliminating trainees. At the same time, m_{ij} for all tasks was sufficiently high so that a_i could have its full effect; the high value of m_{ij} can be inferred from the fact that the candidates were trying for a much-prized opportunity to study a foreign language in a civilian setting and thus escape some of the routines of military service. Finally, one may judge that p_j, the adequacy of presentation, was uniformly high. The instruction was highly systematized and the instructors had had much experience in administering the trial course.

Very similar conditions obtained in the FSI language courses, another situation in which language aptitude test validities were extremely high. To be sure, the course was much longer than the four-day trial course, but in view of the large amount of material to be learned, it is probably the case that o_j for each subtask was relatively small. Further, the courses themselves had been under development for some time and were carefully planned; motivation was probably high in the sense that students were usually willing to put as much time as necessary into the task of learning.

The course conducted by the U. S. Marine Corps during World War II and studied by Williams and Leavitt (1947) is possibly an instance of a situation where both o_j and p_j were quite small. The former condition (small o_j) is judged from the fact that the course is described as exceedingly rigorous—"requiring acquisition of a speaking and reading mastery of Japanese in six months," according to Williams and Leavitt, and yielding an attrition rate of something like 80%. The latter condition (small p_j) is judged from the fact that as far as can be known, the course in Japanese conducted by the U. S. Marine Corps during the war years was not much influenced by the methodological notions which were being developed by linguistic scientists working with the Army. It is quite likely that the learning content was presented by means of fairly traditional grammar-translation methods which put a premium on the students' ability to understand the learning task itself. If this analysis is correct, it is not surprising that the best predictors of success were

measures of verbal intelligence (including the Army Language Aptitude Test, which is chiefly a measure of verbal intelligence).

The very moderately satisfactory validities obtained at the U. S. Military Academy may be interpreted by assuming that o_j (opportunity) was in general much greater (longer) than in the very intensive courses of the U. S. Air Force and the FSI, for the service academy courses resemble, if anything, language courses in civilian universities. This allowed m_{ij} to play a somewhat larger role than it usually plays in intensive courses. As a matter of fact, an attempt was made to measure a variety of academic motivation at West Point. In a questionnaire given at the time the tests were given, each student was asked to classify himself into one or the other of the following categories:

> "A. The type of student who in every subject works for the highest level of accomplishment he can achieve, regardless of whether he thinks the subject valuable for him. Even when the going gets rough, or even when the learning seems unproductive, this student maintains a high level of effort.
>
> "B. The student who works hardest only on the subjects that interest him or that he thinks valuable to him in some way. This type of student is satisfied to get average or even below-average grades in subjects which do not interest him particularly."

The inclusion of this question was prompted by the then-recent finding of Frederiksen and Melville (1954) that interest tests predicted grades in engineering school better for "non-compulsive" students than for "compulsive" students; in contrast to the indirect measures of compulsiveness used by Frederiksen and Melville, the attempt was made here to get at this by a direct question. It seemed that if interest tests predicted grades better for "non-compulsive" (Type B) students, aptitude tests would predict better for "compulsive" (Type A) students, who would at least be highly motivated in the sense of being more willing to spend the amount of time needed to master the material. Table 4.12 presents the comparative validity coefficients for A and B type students, for the four tests which are most similar to the tests in the final battery, as well as for the WPAT Language and Total Score. Although the differences are not statistically significant beyond the 5% level, the results tend to contradict the hypothesis, since the correlations are uniformly higher for Type B students for all tests listed except WPAT Total. Perhaps the characteristic of Type B students is not that they work on what interests them most but that they work chiefly on what they find easy, i.e., have highest aptitude for; this would ac-

Table 4.12

Comparative Validity Coefficients for Selected Tests, "Type A" (Compulsive) vs. "Type B" (Non-Compulsive) Students at the U. S. Military Academy

Test	Criteria (Language Grades)							
	First-Term Grades Adjusted for Previous Training				Two-Year Cumulative Grades			
	Type A		Type B		Type A		Type B	
	r	*N*	*r*	*N*	*r*	*N*	*r*	*N*
Artificial Language Numbers	.28	158	.28	119	.24	158	.40	119
Turse Phonetic Association	.35	137	.48	105	.40	137	.45	105
Words in Sentences	.14	158	.19	119	.19	158	.30	119
Paired Associates	.19	151	.20	137	.20	151	.19	137
WPAT Language Score	.26	305	.25	249	.30	305	.35	249
WPAT Total	.35	305	.29	249	.38	305	.34	249

count for the higher validities of aptitude tests in this group. This interpretation would suggest that in applying the model, m_i should in some circumstances be made to depend to some extent on a_i.

The Nature of Foreign Language Aptitude

Having identified situations in which aptitude, a_{ij}, can be shown to exist as a variable (or complex of variables), there remains the task of describing the nature of foreign language aptitude as it can best be discerned by inspection of the language aptitude tests and consideration of results, including factor analytic results (Carroll, 1958). Accordingly, language aptitude can be considered under the following four headings:

Phonetic Coding. One of the most important abilities required in learning a foreign language is *the ability to "code" auditory phonetic material* in such a way that this material can be recognized, identified, and remembered over something longer than a few seconds. The "coding" is presumably a cognitive process which cannot be directly observed, but something of this sort may be inferred from the following case report: A woman who had received a low score on the Phonetic Script Test was presented with two spoken nonsense syllables /θej; θaej/, and then ten seconds of mental arithmetic to do, after which she was asked to repeat the two syllables. She could not do this, although this task was known to be relatively easy for most people. Further, the woman could herself repeat the syllables accurately when allowed to do it immediately after original presentation. Thus, this ability is not the

ability to make an echoic response to phonetic material, but the ability somehow to "code" or represent it in imagery so that it can be recognized or reproduced after an intervening period filled with other activity. This ability, it would seem, is measured chiefly by the Phonetic Script Test, in which the individual has to learn how a series of speech sounds are represented by alphabetic characters; in order to do this, however, the sounds themselves have to be "coded" or "stored" long enough to be compared with other sounds, and the individual has to build up a considerable repertoire of responses. This ability may also be drawn upon, however, by paired-associates tests utilizing nonsense syllables or paralogs. To a slight extent, also, it may be involved in the Artificial Language Numbers Test, although in this test the individual has considerable opportunity to consolidate his learning of the nonsense materials. It is also measured by the Spelling Clues Test insofar as this represents phonetic-orthographic habits which the individual has learned.

In learning a foreign language, a person low in this ability will have trouble not only in remembering phonetic material, words, forms, etc., but also in mimicking speech sounds. Apparently the process of making an echoic response involves some degree of "phonetic coding," or perhaps it would be better to say phonemic coding because the individual will impose upon his repetition of a heard utterance whatever system of phonemes he has acquired most strongly.

Grammatical Sensitivity. A second important variable in language aptitude is *the ability to handle "grammar,"* i.e., the forms of language and their arrangements in natural utterances. This implies that the individual is sensitive to the functions of words in a variety of contexts. This may be a learned trait, but it is conceivable that variations in this ability may be observed even when the individual has no formal training in grammar. It is postulated that this trait is particularly well measured by the Words in Sentences subtest of the Modern Language Aptitude Battery.

Rote Memory for Foreign Language Materials. A third important variable is that of *rote memorization ability for foreign language materials.* This ability is to be regarded as independent of and different from the phonetic coding variable described above; it has to do with the capacity to learn a large number of these associations in a relatively short time. Though a certain degree of phonetic coding ability is necessary, perhaps prerequisite, those who have requisite phonetic ability may still not be able to hear and remember the relationships. It is

postulated that the Paired Associates Test measures this ability fairly accurately; it is also tapped by the Number Learning Test.

Inductive Language Learning Ability. A fourth variable is what may be called "*inductive language learning ability*." This is the ability to infer linguistic forms, rules, and patterns from new linguistic content itself with a minimum of supervision or guidance. It is not measured to any appreciable degree by the tests of the present final MLAT battery, but it had turned up in certain earlier studies (Carroll, 1958).

The above four factors do not include what is ordinarily called the verbal or verbal knowledge factor, which according to the results of the studies reported is not very important in predicting success. Vocabulary tests do not serve as particularly good predictors, at least in situations where other tests serve well, since the first stages of learning language do not require one to acquire a large vocabulary. On the other hand, the present Spelling Clues Test functions in part as a vocabulary test.

Travers (1954) has attempted to classify aptitude variables into the following categories:

1. Measurement of the extent to which the individual has already acquired the responses required in training.
2. Measurement of the extent to which prerequisite responses have been learned.
3. Measurement of the extent to which "related" responses which facilitate learning have been acquired.
4. Measurement of the ability to make the discriminations necessary to profit from learning.
5. Measurement of motivational variables (anxiety, exploratory drive, etc.).

While it is evident that Travers has tried to stick closely to a kind of parsimonious model which stresses stimulus and response, this framework is not completely satisfactory for schematizing the variables that have been postulated in foreign language aptitude. A sixth category is therefore proposed:

6. Measurement of the extent to which the individual can perform tasks with a behavioral structure characteristic of those required to be learned in the training.

Space will not permit a full explication of what is meant by "behavioral structure"; what is implied is that one cannot describe a task solely in terms of stimuli or in terms of responses, as Travers attempts to do; rather, the tasks must be described in terms of an interaction of stimulus, response, and time variables.

Of the four chief elements which have been postulated in language learning, only one, grammatical sensitivity, could be reasonably regarded as falling under Travers' categories, and even in this case it is debatable whether it might better fall under his category 2, 3, or 4, if not in the new category 6. Further research will be needed to help decide this issue. In the meantime, the remainder are rather clearly to be considered under category 6. For example, phonetic coding is represented by tasks in which it is necessary for the examinee to identify a particular kind of (phonetic) stimulus, associate it with another stimulus which constitutes its symbol or "code" (either overtly or covertly), and demonstrate mastery of this association even after interference from other intervening tasks. Rote memorization involves a somewhat similar behavioral structure, except that the stimulus itself may not involve associations, and characteristically there are a considerable number of associations to be learned in a given time. Inductive language learning, finally, is a matter of how rapidly the subject can utilize a range of contrasting stimulus materials in order to arrive at certain rules of procedure in his future behavior.

APPENDIX

Descriptions of Tests

Anagrams. The task is to write as many words as possible using the letters in the word "occupation"; four minutes.

Artificial Language Learning. This test uses "Tem-Tem," a specially constructed artificial language. From simultaneous presentation of still pictures projected on a screen and spoken Tem-Tem equivalents recorded on tape, S (the Subject) inductively learns to understand Tem-Tem sentences. His learning is then tested by asking him to select pictures corresponding to spoken sentences. The test contains three parts, each representing a lesson with an associated test of 10 items. Total time required: 26 minutes.

Artificial Language Numbers. By tape recording, S is taught a simple artificial system of number expression utilizing nonsense syllables. He is then asked to write down the Arabic numeral equivalents of a list of two- and three-digit numbers in the artificial system, spoken at a fairly rapid pace on the tape. This test utilized only the digits 0, 1, 2, and 3. Total time: 11.5 minutes.

Cooperative Vocabulary Test (Form Z). A non-speeded, wide-range vocabulary test of the conventional multiple-choice type, constructed by Davis and Davis and published in 1949 by the Cooperative Test Division of the Educa-

tional Testing Service. A time limit of 19 minutes was used; a scaled score is computed by procedures prescribed in the test manual.

Devanagari Script. By means of a tape recording lasting about 15 minutes, and a printed worksheet, S is taught the sounds of seven symbols (four consonants and three vowels) in the Devanagari script, the system of writing used in Sanskrit and other languages of India. The test consists of 24 items, in each of which S must indicate which of two trisyllabic words in Devanagari script (the "neither" response also being offered) is spoken on the tape.

Disarranged Letters. Words are given with letters disarranged, e.g., uckd (=duck). The task is to write the word correctly. For each group of items a class name is given; "Birds," "Furniture," etc. Speeded; 42 items, four minutes.

Disarranged Words. S is required to rearrange segments of two words (always an adjective and a noun, respectively) into meaningful order, indicating his solution by marking the number of the last syllable in the second word. Sample:

1	2	3	4	5	
ing	able	dark	notice	en	(Answer: 1)

Score is number right. 40 items, six minutes

Letter-Star Test. S is presented with patterns of letters and asterisks such as * Y * S, and is to invent a meaningful phrase fitting this pattern by substituting a word for each symbol, with the restriction that words substituted for capital letters must begin with the letter indicated. Score is the number of items completed in five minutes; 50 items.

Number Learning. A slightly harder version of Artificial Language Numbers utilizing the digits 0, 1, 2, 3, and 4. Total time: ten minutes. Constitutes Part 1 of the Psi-Lambda Foreign Language Aptitude Battery.

Oriental Script. Similar to Devanagari Script except that this test is based on the system of writing used in classical Mongolian. 50 items; the tape takes 15 minutes.

Paired Associates. Examinee studies a list of 24 "Turkish-English" vocabulary equivalents for two minutes; in the next two minutes, he practices recalling the English meanings, and in the final four minutes he completes from memory a multiple-choice test of the presented vocabulary. Constitutes Part 5 of the Psi-Lambda Foreign Language Aptitude Battery.

Perdašeb. "Perdašeb" is an artificial language constructed specially for this test (designed by S. M. Sapon), which is an attempt to duplicate, in miniature, the learning situation in the typical "grammar-translation" foreign language course. The student is taught the grammer of Perdašeb in a series of taped

lessons, during which he follows material in the test booklet as in a textbook. He then checks the accuracy of translations from English to Perdašeb and from Perdašeb to English. 26 items; 38 minutes.

Phonetic Discrimination I, Test P-120. Designed to measure the ability to discriminate between minimally different speech sounds in various foreign languages. Each item consists of three spoken quasi-words presented auditorily on a tape; two of these are precisely the same, while the third differs in its medial sound. S is to indicate which of the three stimuli is different. 50 items.

Phonetic Discrimination II. Similar to Phonetic Discrimination I except that the items follow a multiple-choice pattern: a syllable is heard, after which S is to identify which of four syllables heard subsequently is the same as the first. 25 items.

Phonetic Script. S learns a series of phonetic symbols for some of the phonemes of English by listening to pronunciations recorded on magnetic tape and following syllables printed in phonetic symbols on the test paper; after every five items for the first 30 items S goes back and gets a test on the material just learned. After the 30-item learning period, there is a 30-item test in which S must indicate, for each item, which of four phonemically-printed syllables is pronounced on the tape. All the phonemes used in the test are in English, and no fine phonetic discrimination is required. 60 items; 15 minutes. (The first 30 items may be separately scored, and constitute Part 2 of the Psi-Lambda Foreign Language Aptitude Battery.)

Phrase Completion. S is given a number of incomplete sentences or phrases, such as "But it's all ----," which are to be completed with the first word that comes to mind. Scoring based on community of response, high score for most frequent responses. 24 items; no time limit.

Picture Naming. S writes the first letters of the names of a series of common objects as pictured. Highly speeded: 147 items, two minutes.

Rhyming. Subjects are to give as many rhymes as possible to each of four words (low, case, speak, lose), one minute allowed per word.

Same-Opposites (F-S version). A multiple-choice test similar to an ordinary vocabulary test, except that in each item S must find either a synonym or an antonym of the given word depending on whether it is marked with an asterisk. S presumably must shift set rapidly and often. Scored for both accuracy and speed. 100 items; four minutes.

Spelling Clues. An adaptation of the Turse Phonetic Association Test to objective scoring. S chooses which of five words has the same meaning as the

word represented in abbreviated form. Sample: kataklzm = 1. mountain lion; 2. disaster; 3. sheep; 4. chemical reagent; 5. population. Highly speeded: 50 items, five to eight minutes. Constitutes Part 3 of Psi-Lambda Foreign Language Aptitude Test.

Turse Phonetic Association. This is Test 3 of the Turse Shorthand Aptitude Test published by the World Book Company. S must spell out correctly, in writing, a word which is printed in abbreviated form "approximately as it is pronounced." Sample: *tox* = "talks." 60 items; five minutes.

Turse Spelling. This is Test 2 of the Turse Shorthand Aptitude Test published by the World Book Company. S is asked to identify which, if any, of three alternative spellings of a word is correct. 45 items; four minutes.

Turse Word Discrimination. This is Test 5 of the Turse Shorthand Aptitude Test published by the World Book Company. Choose, from among words likely to be confused, the word that fits correctly in a sentence. Sample: The (leak lick lack lake) of water made irrigation necessary. Speeded: 30 items, five minutes.

Verbal Enumeration (F-S version). This test was designed by S. M. Sapon to measure a postulated "flexibility of set" factor which might be relevant in switching rapidly from one language to another. Use is made of Thurstone's verbal enumeration test format, in which S marks all words in a column which are names of things in the category given by the column heading. In this version, however, columns contain many words falling in classifications used in preceding columns, and S must presumably avoid carrying over the set established in the preceding columns. Scored for both speed and accuracy. 18 columns of 40 words each.

Word Elements. S is required to select, inductively, examples of Latin and Greek roots and affixes found in English words and to give their meanings in a multiple-choice test. 30 items, 11 minutes.

Words in Sentences. Designed to measure ability to understand the function of words and phrases in sentence structure, without calling upon knowledge of grammatical terminology. Each item consists of a key sentence with a word or phrase printed in capital letters, followed by one or more sentences with words and phrases underlined and numbered. S is directed to pick the word or phrase in the second sentence or sentence group which does the same thing in that sentence as the capitalized word does in the key sentence. Sample:

He spoke VERY well of you.

Suddenly the music became *quite loud.*

1 2 3 4

30 items; 16 minutes (allowing most to finish). Constitutes Part 4 of the Psi-Lambda Foreign Language Aptitude Battery.

Word Squares. For 15 seconds S views a word square composed of sets of homophones, such as "cent," "scent," "sent," after which he selects which of a number of word squares is exactly the same as the one he has viewed. This is done solely from memory. The test contains three such word squares.

REFERENCES

Bottke, K. G. & Milligan, E. E. Test of aural and oral aptitude for foreign language study. *Modern Language J.*, 1945, *29*, 705-709.

Carroll, J. B. A factor analysis of verbal abilities. *Psychometrika,* 1941, *6,* 279-307.

Carroll, J. B. Some principles of language testing. *Georgetown Univ. monogr. series in languages and linguistics,* 1953, No. 4, 6-10.

Carroll, J. B. A factor analysis of two foreign language aptitude batteries. *J. gen. Psychol.,* 1958, *59,* 3-19.

Carroll, J. B. Research on teaching foreign languages. In N. L. Gage (Ed.), *Handbook of research on teaching.* 1962. (In press)

Carroll, J. B. & Sapon, S. M. *Modern Language Aptitude Test, Form A.* Test Booklet, Practice Exercise Sheet, Answer Sheet, Answer Keys, Manual, Tape. New York: The Psychological Corporation, 1958, 1959.

Dorcus, R. M., Mount, G. E., & Jones, Margaret H. *Construction and validation of foreign language aptitude tests.* Los Angeles: Department of Psychology, University of California, 1952. (Personnel Research Branch Research Report 993, The Adjutant General's Office, Department of the Army, Contract DA-49-083 OSA-75 PR 3576.)

Frederiksen, N. & Melville, S. D. Differential predictability in the use of test scores. *Educ. psychol. Measmt,* 1954, *14,* 647-656.

French, J. W. *The description of aptitude and achievement tests in terms of rotated factors.* Chicago: University of Chicago Press, 1951. Psychometr. Monogr., No. 5)

Frith, J. R. Selection for language training by a trial course. *Georgetown Univ. monogr. series in languages and linguistics,* 1953, No. 4, 10-15.

Harding, F. D., Jr. *Language aptitude tests as predictors of success in a trial Russian course.* Lackland AFB, Personnel Laboratory, Air Force Personnel and Training Research Center, 1956. Technical Memorandum PL-TM-56-5. (a)

Harding, F. D., Jr. *Language aptitude tests as predictors of success in a trial Mandarin Chinese course.* Lackland AFB, Personnel Laboratory, Air Force Personnel and Training Research Center, 1956. Technical Memorandum PL-TM-56-8. (b)

Harding, F. D., Jr. Tests as selectors of language students. *Modern Language J.*, 1958, *42*, 120-122.

Harding, F. D., Jr. & McWilliams, J. T. *Language aptitude tests as predictors of success in a six-month Russian course.* Lackland AFB, Personnel Laboratory, Air Force Personnel and Training Research Center, 1957. AFPTRC-TN-57-86.

Henmon, V. A. C., Bohan, J. E. & Brigham, C. C. *Prognosis tests in the modern foreign languages.* New York: Macmillan, 1929.

Hunt, Thelma, Wallace, F. C., Doran, S., Buynitzky, K. C. & Schwarz, R. E. *George Washington University Language Aptitude Test.* Washington: Center for Psychological Services, 1929.

Kopstein, F. F. & Roshal, S. M. Learning foreign vocabulary from pictures vs. words. *Amer. Psychologist*, 1954, *9*, 407-408.

Luria, M. A. & Orleans, J. S. *Luria-Orleans Modern Language Prognosis Test.* Yonkers: World Book Co., 1928, 1930.

McBee, G. & Duke, R. L. Relationship between intelligence, scholastic motivation, and academic achievement. *Psychol. Rep.*, 1960, *6*, 3-8.

Sapon, S. M. A work-sample test for foreign language prognosis. *J. Psychol.*, 1955, *39*, 97-104.

Sarason, S. B. & Mandler, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, *47*, 810-817.

Stoddard, G. D. & VanderBeke, G. E. *Iowa Placement Examinations*, Series FAI, Revised, A. Iowa City Extension Division, State University of Iowa, 1925.

Symonds, P. M. *Foreign language prognosis test.* New York: Teachers College, Bureau of Publications, 1930.

Travers, R. M. W. *An inquiry into the problem of predicting achievement.* Lackland AFB, Personnel Research Laboratory, Air Force Personnel and Training Research Center, 1954. AFPTRC-TR-54-93.

Travers, R. M. W. Personnel selection and classification research as a laboratory science. *Educ. psychol. Measmt*, 1956, *16*, 195-208.

Williams, S. B. & Leavitt, H. J. Prediction of success in learning Japanese. *J. appl. Psychol.*, 1947, *31*, 164-168.